1995

The Effects of Assessor and Assessee Gender, Ethnicity, and Assessor's Role on Performance Assessment of Teachers.

Shana Lewitt Schuyten
Louisiana State University and Agricultural & Mechanical College

INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

UMI

A Bell & Howell Information Company
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
313/761-4700 800/521-0600

THE EFFECTS OF ASSESSOR AND ASSESSEE GENDER, ETHNICITY, AND ASSESSOR'S ROLE ON PERFORMANCE ASSESSMENT OF TEACHERS

A Dissertation

Submitted to the Graduate Faculty of the
Louisiana State University and
Agricultural and Mechanical College
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

in

The Department of Administrative and Foundational Services

by
Shana Lewitt Schuyten
B.A., Tulane University, 1991
August, 1995

ACKNOWLEDGEMENTS

Before I can begin to thank all of the many wonderful people who have contributed so much and supported me for so long, I must first give praise to my higher power. Without the Lord almighty's grace I would have never have made it to where I am. In fact, I may never have known where I was going. The Lord has touched my life in so many ways, giving me the strength to go on when I had none. He has been right beside me to witness my successes as well as my failures, but it was he who gave me the courage to keep trying and to never give up on my dream.

Another person who has been my rock (and so the Greek derivative of his name means so) is my husband and best friend, Peter. Peter has always believed in me, continued to support me, and always loved me. He has sacrificed a great deal that I might pursue what has been so important to me. My gratefulness and admiration are eternal, Thank you, Peter.

I wish to thank all my family and friends who have been my saving grace. My family, both immediate and extended, has been so patient, so understanding and so supportive. A special thanks to a dear friend, Catherine Glascock. Catherine has been both a motivator and guiding light for me. I wish to personally thank my aunt and uncle, Dr. Chad and Alberta Ellett, who challenged me from an early age to be the best that I could be. They were the first of many to instill

in me the values of personal commitment, dedication, and sacrifice. These very same values have made me a successful doctoral student, and I am truly grateful for these lifelong lessons.

I would like to express my sincere appreciation to all of my committee members who have been so supportive of me over the last four years. Their sincere friendship, as well as enduring professionalism has meant more to me than they will ever know. When I thought that I had nothing more to argue about, I could always go to Dr. Abbas Tashakkori, my teacher, advisor, friend, and debate partner. As my major advisor, Dr. Tashakkori has always been very willing and supportive of my efforts. Thanks to Dr. Tashakkori, I have never stopped learning nor have I wanted to. When I thought that I could no longer stand to look at one more statistical formula, it was Dr. Charles Teddlie who helped me regain my sanity. I thank him for breaking up the quantitative monotony! I also thank him for his enduring support and encouragement. When I just needed to hear that I wasn't as confused as I sounded, Dr. Eugene Kennedy was always very comforting and soothing. Dr. Kennedy has been a great security blanket although I doubt he knows it. When I just needed to remind myself to stop and smell the roses, I thought of Dr. Diane Taylor. Dr. Taylor has been a tremendous role model for me, reminding me that women can excel in the higher education field.

Truly, you have all been an inspiration to me. Your unwavering support, guidance, and encouragement have been among my most memorable experiences. I thank you all for helping me to realize this awesome dream, as I could never have done it without you.

TABLE OF CONTENTS

LIST OF TABLES

ABSTRACT

Two studies were conducted as part of this research effort. The purpose of these studies was to determine the effects of assessor and assessee gender, ethnicity, and assessment role on performance observation ratings. Study I was Causal Comparative in nature and involved analyzing actual performance observation ratings received on the Louisiana Teacher Assessment Instruments (LTAI). Study II was an experimental study and involved analyzing experimentally manipulated, teacher-performance-observation ratings received on an instrument entitled "Survey of Effective Teaching Behavior." The data were collected in the spring of 1995 and expand upon the findings of Study I.

There were essentially three different issues of bias to be addressed. Issue I addressed gender bias, Issue II addressed ethnicity bias, and Issue III addressed role-related bias within the assessment ratings, that is, Issue III examined the prevalence of bias attributable to the "type of "role" assumed by the assessors within the assessment context.

Study I results indicate significant main effects of assessee gender. Female assessees scored higher than male assessees on all components. The results also indicate that some differences in assessment ratings are attributable to assessee/assessee ethnicity. Caucasian assessees had consistently higher ratings than African-American assessees,

regardless of assessor ethnicity. Results regarding Role Bias indicated that only in one of the eight components are the differences in assessment ratings attributable to assessor role. In addition, those assessors assuming the role of principal give higher ratings, yet master teachers have a slightly higher overall mean component rating.

It is concluded that despite some statistically significant effects, magnitude of bias due to gender, ethnicity, or role was small. However, it is methodologically important that we examine the possibility of gender, ethnicity, or role biases that may devalue assessment results. As the nation moves toward teacher assessment systems that rely on observational rating performances, we must be prepared to extrapolate true assessment ratings from those that are confounded by bias. Differences in assessment results are tolerable but only if they are not the result of gender, ethnicity, or role biases rather than due to true differences in assessees' performance.

CHAPTER 1: INTRODUCTION

The teaching profession is in the midst of dramatic reform. The impetus for this reform is the growing public discontent over the quality of education in the nation's schools (Tanner, 1993). As part of the reform, efforts are undertaken to ensure that there is a high quality of education in the schools across the country. Widespread changes are being proposed regarding the ways teachers are educated, trained, evaluated and certified. At the forefront of these changes are programs of induction and evaluation for beginning teachers. These programs were developed for the enhancement and improvement of teaching in schools.

The attention to beginning teachers can be attributed to research over the past two decades (e.g., Ryan, 1979; Tisher, 1978) that has found that the first year of teaching is critical and is often a difficult transition period in teacher development (Hoffman et al, 1986). In a comprehensive study of beginning teaching, McDonald (1980) reported:

> For most teachers, the initial experiences of teaching are traumatic events out of which they emerge defeated, depressed, constrained or with a sense of efficacy, confidence and growing sureness in teaching skills.(p.5)

McDonald also speaks of beginning teachers as "abandoned" by the institutions where they received their preservice training. They also are considered "peers" to all other

teachers and by their employers. They have traditionally been left to their own devices to endure the first few years of teaching alone.

The reform movement, as it relates to beginning teachers, is implemented primarily through policy initiatives at the state level. Few local school district policy makers and administrators are given the responsibility of devising and implementing methods of teacher evaluation. In 1980, there were only five district-supported programs for beginning teachers, of which two were at developmental stages (Hoffman et al, 1986). A more recent survey indicated that more than 65% of all school districts in the United States have instituted some type of standardized teacher appraisal system (Katims & Henderson, 1990).

At the time of Hoffman's study (1980), only one state, Georgia, was active in the area of induction and evaluation of beginning teachers; since that time numerous states have become active in this arena. Eight years ago a national survey of state activity in programs for beginning teachers identified 18 states with programs in advanced planning stages and 4 states with operational programs (Defino & Hoffman, 1986). Among those states were Georgia, Florida, Connecticut, Arizona, and Texas. All have mandated large-scale standardized teacher-performance-appraisal systems as a part of efforts to reform and improve education in those states (Greenfield, 1987).

To investigate the phenomenon of Competency Assessment, Sandefur (1983) conducted four annual surveys of the 50 states to provide data for analyzing nationwide trends in competency assessment of teachers. Ten years ago, the findings indicated that most state plans for teacher-competency assessment included testing one or more areas of basic skills, professional or pedagogical skills, and academic knowledge. The testing took place at the entry level, admission to the teacher-education program, or prior to certification. At that time, a growing number of states had began to require an internship or beginning teacher year with adequate assessment before initial certification was awarded. Sandefur's data analysis found that state competency assessment of programs grew rapidly over the prior six years and will continue to increase. He also indicated that continuing trends emphasizing testing in the basic skill areas will be used for certification purposes. Data also indicated that fewer states were using legislative action to mandate competency assessment of teachers; instead more states were relying on the regulations of their state departments of education.

Currently, there is an increased demand for the identification of competent teachers within school systems across the nation. This demand, coupled with the availability of research and assessment instruments, led to the development of large-scale teacher-assessment systems which

were legislatively enacted in such states as Arkansas, Connecticut, Florida, Georgia, Kentucky, Mississippi, Missouri, North Carolina, South Carolina, Tennessee, Texas, and Virginia (Chauvin & Ellett, 1991; Ellett 1990). As many as 18 states utilized evaluation systems that were designed to include "on the job" assessment for purposes of teacher certification, merit pay, career ladders, and professional development (Association of Teacher Educators, 1988; Chauvin et al., 1991; United States Department of Education, 1987). If this trend continues, other states will follow their lead.

At the heart of these beginning teacher-evaluation programs are classroom-observation systems designed for certification or employment decisions. Any system relying on observation as a mechanism for rating performance might be affected by the limits inherent in observational methods. Observational methods are advantageous due to the wealth of descriptive information they provide, but there are some pronounced limitations.

Basic Terminology

Before the limitations of observation can be discussed, a review of associated terminology from the research literature must be discussed. The term observation refers to a form of data collection which results in detailed descriptions of people's activities, behaviors, actions and the full range of interpersonal interactions that are part of observable human experience (Patton, 1990). These detailed

descriptions can be the result of "systematic" observation, in which a trained observer employs a predefined observation form (Borg & Gall, 1989). The observer must remain alert and attentive, constantly noticing and perceiving activity as it occurs. In most educational research textbooks and journals, the term "observer" is often used interchangeably with the term "rater." When used in conjunction with activities deemed as evaluation, appraisal, or assessment, the observer or rater is more appropriately called "evaluator, appraiser, and assessor." When these observers or raters make errors in attribution or perception, the errors are called "observer errors" and "rater errors." Since these terms are used interchangeably and in an overlapping manner, the literature can be confusing at times. Unfortunately, much of the literature continues to promote the confusion by using the overlapping terminology.

Despite all of this, while we know what an observer is and does, an observer is often called a rater. Observers do not necessarily rate the behaviors they observe, although they could. Raters do not necessarily observe the behaviors that they rate, although they also could.

The terms evaluation and assessment are often synonymously used in a variety of ways. Evaluation combines measures with other information to establish the desirability and importance of what we have observed (Oosterhoof, 1990). It is thought of as a professional judgement or process which

allows one to make a judgement about the desirability or value of something (Mehrens & Lehmann, 1991). However, a second popular concept of evaluation interprets it as the determination of the congruence between performance and objectives (Mehrens & Lehmann, 1991). Assessment is also used broadly and indicates the use of formal and informal data=gathering procedures and the combining of the data in a global fashion to reach an overall judgement (Mehrens & Lehmann, 1991).

With the understanding that the terms evaluation and assessment are synonymous as well, it is only plausible that the terms evaluator and assessor are as well. The term rater however, is not synonymous with the terms evaluator and assessor. It is essential to make the distinction that a rater uses measurement to quantify the characteristics of observation. Raters rely on a form of measurement to determine the desirability and value of what is observed. Evaluators and assessors are not limited to measurement approaches such as ratings to make judgements about the importance, value, or desirability of their observations.

Bias

Bias is generally defined as systematic error in measurement and/or observation and it is the most obvious and limiting phenomenon associated with observational techniques. Bias is an inclination or preference, especially one that interferes with impartial judgement; in effect, bias is

prejudice (Webster, 1988). Bias is a naturally occurring phenomena which is exercised consciously in overt manners as well as unconsciously in covert manners.

Gender Bias

Gender Bias occurs when the sex (male or female) of rater and/or ratee interferes with the rater's ability to conduct objective observations. The bias is reflected in the ratings given and is directly attributable to the gender of one or more of the individuals involved. It is likely that the observer incorporates his or her attitudes and perceptions about a particular sex into observation, projecting them into the ratings that they give to the individual.

In a review of existing literature between 1932 and 1979, Feldman (1983) found that females were consistently rated higher than males. Female subjects also rated performance consistently higher than did male subjects (at least on some items) in studies by Basow and Distenfeld (1985), Basow and Howe (1987), Bennett (1982), and Harris (1976). In a 1989 study which investigated the effects of gender, status, and effective teaching on the evaluation of college instruction, there was also some evidence of gender bias (Dukes & Victoria, 1989). In this study, male subjects rated female professors higher than female subjects. In all of these studies, the sex of the rater and ratee interfered

with the objectivity and credibility of the rater, as well as the validity of the ratings.

In other gender research, such as Levenson, Bufford, Bonnoe and Davis (1975) and Tieman and Rankin-Ullock (1985), male subjects rated performance higher than female subjects. Methodology developed by Goldberg (1968) was ideally suited to the examination of teaching evaluations. The basic mechanism of the methodology in the original experiment was to present subjects with the stimulus material in the form of a journal article that was identical in all respects except for the gender of the author. The procedures were used to "uncover the differential evaluations of males and females for identical performances on written tasks" (Goldberg, 1968). More than 20 years of subsequent research have extended the technique to include topics as diverse as the evaluation of works of art (Etaugh and Sanders 1974; Peterson, Keisler, and Goldberg 1971), the evaluation of helping behavior (Taynor and Deaux 1975), lawyers' performance in a courtroom (Abramson, Goldberg, Greenberg, and Abramson 1977), and admission of a student to graduate school (Tawil and Costello 1983). All the results uncovered a bias in which subjects continually favored a male author over a female. Subsequent research showed that males were evaluated higher than females for the same performance and that the status of the person being evaluated altered the situation. Although it is clear that the literature has produced

findings about the relationship between the rater and ratee's gender, the direction of these effects on the evaluation of effective teaching is not clear.

Ethnicity Bias

Ethnicity Bias occurs when the race (Caucasian, Native American, African American etc.) of rater and/or ratee interferes with the rater's ability to conduct objective observations. The bias is reflected in ratings and is attributable to the ethnicity of one or more of the individuals involved. It is probable that observers incorporate their attitudes, perceptions, and ethnic stereotypes about a particular race into observation and project them into the ratings they give to the individual. Subsequently, these ratings are not valid measures of the individual's performance and are unwarranted.

Research addressing the issues of ethnicity and bias within assessment and evaluation is limited. However, the existing literature suggests that minority teachers face negative effects from evaluations. Minority educators are concerned that evaluations be fair, accurate, and productive (Peterson, Dehyle & Watkins, 1988). In Martocchio and Whitener's (1992) study, fairness in personnel selection was addressed. Martocchio and Whitener (1992) suggest that minorities may be receiving lower performance appraisals due to prejudice or bias on the part of their supervisors. Several researchers have explored the extent to which

performance measures, particularly supervisory ratings, might be biased. Kraiger and Ford (1985) performed a meta-analysis in which Black and White raters evaluated Black and White ratees. They found that raters tend to evaluate ratees of their own race higher than ratees of the opposite race. Kraiger et al. note that since supervisors are predominantly White, this result suggests that minority ratees may be receiving lower evaluations. Pulakos, White, Oppler, and Borman (1989) investigated the effects of rater and ratee race on ratings collected on over 8,000 U.S. Army personnel. While they consistently found significant race effects, these effects explained only a small amount of variance in ratings.

Role Perception Bias

Role Perception Bias occurs when the function that a person serves and the role that he or she assumes within the observational schema affects the ability to observe objectively. Research suggests that the role a person serves within an observational assessment process is critical and must be clearly understood (Ellett & Capie, 1985, Acheson, Smith & Stuart, 1986, Garland, 1989). Literature addressing the many roles and conflicts of persons who conduct observations to evaluate is abundant (Acheson, Smith & Stuart, 1986, Price, 1989, Collins, 1990).

One of the most obvious conflicts is experienced by the building principal, who is expected to be a hard-nosed evaluator of teachers as well as a kind, sympathetic, and

helpful supervisor of instruction (Acheson, Smith & Stuart, 1986). Additional literature suggests that the principal (or immediate supervisor) may experience role conflict in trying to serve as the instructional leader and as an administrative decision maker (Duckett, 1985; Stanley & Popham, 1988). Research addressing on-campus (principals) and off-campus (university faculty members) evaluators of teachers has also been conducted. Findings indicate that on-campus evaluators/principals tend to award higher scores than off-campus evaluators or those who do not directly serve as the teacher's immediate supervisor (Cronin & Capie, 1986; Ellett & Capie, 1985; Ellett, Teddlie & Niak, 1991; Kelly, 1985; Rose & Hutnh, 1984, Wise et al., 1984). Clearly, the persistent problem of performing in both a supervisory and evaluatory capacity remains a dilemma for principals and administrators.

As the literature has illustrated, evaluators experience bias within many disciplines. In the educational arena, bias within performance assessment can not run rampant, nor can it be ignored. Policy makers and key stake holders must ensure that they are utilizing an assessment system that minimizes bias and fosters valid and reliable assessment procedures and personnel. For this reason, both researchers and educators must investigate the possibilities of biases resulting from ethnicity, gender, and role within educational assessment systems.

Statement of Purpose

Two studies were conducted as part of this research. The purpose of these studies was to determine the effects of assessor and assessee gender, ethnicity and assessment role on performance observation ratings. Study I will involved analyzing actual performance-observation ratings received on the Louisiana Teacher Assessment Instruments (LTAI). The data in this study were collected during the 1993-94 Pilot phase of the Louisiana Teacher Assessment Program for Interns. Study II involved analyzing a contrived performance observation rating received on the Survey of Effective Teaching Behavior. The data in this study was collected in the spring of 1995 to follow up and to expand upon the findings of Study I.

There were essentially three different issues of bias addressed within these studies. Issue I addressed gender bias within assessment ratings. Issue II addressed race/ethnic bias within assessment ratings. Issue III addressed role-related bias within the assessment ratings, that is, Issue III examined the prevalence of bias attributable to the type of "role" assumed by the assessors within the assessment context.

Research Questions and Hypotheses

An extensive literature review of research addressing gender, ethnic and role biases supported the following hypotheses. The hypotheses are stated for purposes of both

Study I and Study II. Some of the hypotheses may be directional hypotheses, which will be better understood after reading Chapter Two.

Issue I: Gender Bias

The purpose of Issue I was to examine the data for any evidence of gender bias in the assessment ratings. The independent variable for Issue I is the gender (Male/Female) of the assessor and the assessee, and the dependent variable is the assessee's performance observation ratings.

The research questions pertaining to Issue I were:

1. Do male and female assessors differ in their overall ratings of assessees?
2. Is there a difference in the average ratings of male and female assessees?
3. Does the gender of both the assessor and the assessee interact to influence the performance ratings of teaching effectiveness? (That is, do same-sex evaluation ratings differ from opposite-sex evaluation ratings?)

The Hypotheses pertaining to Issue I were:

1a. Female assessors rate both male and female assessees higher than male assessors will.

2a. There is a statistically significant difference in the mean ratings received by male and female assessees.

Issue II: Ethnicity Bias

The purpose of Issue II was to examine the data for any evidence of ethnic bias in the assessment ratings. The

independent variable for Issue II is the ethnicity of both the assessor and the assessee, and the dependent variable is the assessee's performance observation ratings. For purposes of this study, the independent variable of ethnicity was limited to 2 levels: African American and Caucasian.

The research questions pertaining to Issue II were:

4. Do African American and Caucasian assessors differ in their average ratings of assessees?

5. Is there a difference in the average ratings of African American and Caucasian Assessees?

6. Does the ethnicity of both the assessor and the assessee interact to influence the performance ratings of teacher effectiveness?(That is, do same-ethnicity evaluation ratings differ from opposite-ethnicity evaluation ratings?)

The Hypotheses pertaining to Issue II were:

4a. There is a statistically significant difference between the mean ratings given by African American and Caucasian assessors as measured by the LTAI.

5a. There is a statistically significant difference in the mean ratings received by African American and Caucasian assessees.

Issue III: Role Perception Bias

The purpose of Issue III is to examine the data for any evidence of assessor role bias in the assessment ratings. The independent variable for Issue III is the role that assessor serves in the assessment process, (e.g., principal or

external assessors), and the dependent variable is the assessee's performance-observation ratings. For purposes of this study, the independent variable entitled "role" is actually addressing the "type" of assessor. There are three types (or roles) of assessors within the assessment process: principal, master teacher, and external assessors (i.e., university faculty member).

The research questions pertaining to Issue III were:

7. Is there a difference in average ratings given by the three types of assessors (principal, master teacher, and external assessors)?

8. If there is a difference in average ratings given by the three types of assessors, do principals give higher average ratings than principals and external assessors?

The Hypotheses pertaining to Issue III are:

7a. There is a statistically significant difference between the mean ratings given by the three types of assessors (principal, master teacher, and external assessors) as measured by the LTAI.

8a. Principals are more lenient in rating assessees than other raters. Therefore, principals will have the highest mean ratings among the three types of assessors.

Background and Definition of Variables

Following the example of other states, Louisiana has a legislative mandate to evaluate all beginning teachers (interns) in public schools. The program, *Louisiana Teacher*

Assessment Program for Interns, has two basic uses: 1) to develop information about the intern teacher's competence that can be used to structure instructional improvement activities and 2) to develop information upon which sound decisions about the intern teacher's qualifications for certification can be based (Louisiana Department of Education [LDE], 1993).

To serve these purposes, the system is used during the first semester of an intern teacher's employment to develop a profile of strengths and needs. It is used by the educational support team to assist intern teachers in their professional growth and development. This is considered the **formative** (support semester) part of the evaluation program.

During the second semester of the intern teacher's employment, the system is used to collect data used by the team to recommend either certification or continuation for a second year in the induction/intern program. This is considered the **summative** (assessment semester) part of the evaluation program. If an intern teacher is recommended for a second year in the induction program, assessment and assistance are continued in accordance with the pattern used in year one. An intern who does not demonstrate competence by the end of the second year will be denied certification (LDE, 1993).

The Louisiana Teacher Assessment Program for Interns (LTAPI) is composed of two primary data collection methods:

classroom observation and structured interview. The classroom observation and interview processes utilize instruments and procedures designed to collect data that are directly related to the **Louisiana Components of Effective Teaching (LCET)**.

The LCET is best described as a three-faceted tier of both skills and knowledge which was defined by a team of educators as essential to successful instruction. Schwab (1991) refers to this type of teacher evaluation as being "research based," involving pertinent classroom observations, focusing on indicators that show correlation with teaching success and not relying on global impressions or idiosyncratic variables.

These LCET indicators are hierarchal in nature. The top level of the hierarchy of skills and knowledge is the **Domain** level. These domains can contain and are defined by one or more components. The middle level of **Components** is defined by the lower level of **Attributes**. Together these three levels of teaching skills and knowledge form the **"Assessment Criteria"** (LDE, 1993) (See APPENDIX A). This research-based approach provides independent indicators of teacher effectiveness. The evaluations of the intern teachers are based on independent item (LCET indicators) scores and not on the premise that all indicators must be present in a particular observation or that they be demonstrated in any sequence or pattern (Schwab, 1991).

For purposes of the LTAI, the concept of a **domain** is defined as a major area of teaching responsibilities. Domains are broad and can be difficult to measure; therefore, additional information is needed about the domain for it to be measured. This is the purpose of components and attributes. A **component** is defined as a critical function within a domain and an **attribute** as a behavior that relates to and helps to define a component (LDE, 1993).

The intern teacher will be assigned to a team of three highly competent, experienced educators **(assessors)**, with each conducting a minimum of one visit to the intern's classroom during each semester of the year. During these visits each assessor will conduct the evaluation utilizing the **Louisiana Teacher Assessment Instrument (LTAI)**. The LTAI pilot test uses a three-point rating scale designed to allow formative feedback to the intern teacher. The three-point rating scale is used in all instruments and to determine ratings on all components and attributes during each individual assessment visit. The three points are defined as 1) Needs improvement, 2) Proficient, and 3) Commendable. As will be discussed in Chapter Three, for the present study, the eight component scores and overall rating were constructed by summing the attribute ratings.

Significance of Studies

Foschi and Lawler's writings (1994) describe performance evaluations as "judgements about the relative success of an

actor at one or more tasks." In effect, a performance evaluation is a type of social perception. As such, it inevitably entails "forming beliefs about the quality of a person's task performance based upon perceptions of the person's activities." When performance evaluations are not structured in such a way that successful and unsuccessful outcomes are distinct and easy to judge, evaluation may be difficult. When this is the case, perceptual biases such as those mentioned may come into play. These can have a powerful effect on performance evaluations. Perceptual biases involving social characteristics such as race, sex, age, and education may ultimately influence the assessment process.

It is important that we examine the possibility of any gender, ethnic, or role biases that may devalue or discredit the significance of assessment results. As the nation moves toward teacher assessment systems that rely on observational rating performances, we must be prepared to extrapolate true assessment ratings from those that are confounded by observer bias. Differences in assessment results are tolerable, but not if they are the result of observer biases rather than true differences in the assessee's performance.

CHAPTER 2: REVIEW OF LITERATURE

Past and Current State Teacher Evaluation Programs

Within the last decade there has been a wave of teacher evaluation initiatives in response to legislation proposed in a number of states. The purpose of this legislation has been to develop and improve upon state-wide teacher certification and appraisal systems.

In 1980, the state of Georgia implemented the first "systematic statewide effort to evaluate on-the-job performance of teachers though the application of the Teacher Performance Assessment Instruments (TPAI) (Capie, Anderson, Johnson & Ellett, 1980) to the initial, professional certification of all beginning teachers through the use of a classroom-based, large-scale teacher evaluation system.

For approximately ten years, the state of Georgia evaluated beginning teachers through observations completed by an external data collector, an administrator, and a teacher. The instrument designed, the Teacher Performance Assessment Instrument (TPAI), contains over 10 competencies and 30 indicators. Approximately three years ago, the state discontinued this program and adopted a new one. Currently the State of Georgia Teacher Evaluation Program is being used to evaluate all teachers, both beginning and veterans. The program is a formal, annual evaluation program that serves to determine if Georgia teachers are performing satisfactorily,

but not for making certification decisions. Ultimately the evaluation program is tied to salary. The Georgia Teacher Evaluation Program has an observation instrument (GTOI) and an additional instrument they call the Georgia Teacher Duties and Responsibilities Instrument (GTDRI), which is scored by exception. Scores are combined on both of these instruments, and teachers are rated as satisfactory or unsatisfactory on the criteria of tasks and dimensions, which are the foundation of the instruments (CDE, 1994).

The state of Texas also employs a state-wide teacher appraisal program. The Texas Teacher Appraisal System (TTAS) was based on early research of the 1970's and early '80's. The instrument was implemented across the state in 1986 and is based primarily on classroom teaching behavior (Barnes, 1987; Texas Education Agency, 1988). The TTAS contains five domains. These can cover 13 criteria, which can be divided into 65 behavioral indicators that have been previously identified and defined (Texas Education Agency, 1988). The first four of the five domains are based on classroom teaching performance, and the fifth domain is related to professional growth and development. Texas requires all teachers, regardless of subject area or grade level taught, to be assessed using the TATS. Separate observations are conducted by a primary appraiser who is normally the teacher's immediate supervisor and a second appraiser, who could possibly be another district administrator. To date,

Texas is the most populous state to enforce statewide appraisal utilizing observational methods (Tyson & Silverman, 1994).

In Ohio, comprehensive studies are underway to observe and assess the most useful method of teacher evaluation in the public school districts. Comparison data with seven theoretical models from the literature on teacher evaluation indicate that more than 84% of Ohio's responding districts use a Traditional Trait Model to evaluate their teachers. In this particular model, teacher evaluation is conducted in terms of traits. Evaluation is grounded in the presence or absence of these traits (Marczely & Bernadette, 1992).

Like many of the southeast states, Louisiana has developed and implemented a statewide teacher assessment program. The Louisiana STAR (System for Teaching and Learning Assessment and Review) was developed over eight years ago in response to two specific legislative mandates. These mandates were the Teaching Internship Law (1984) and 2) the Children's First Act (1988). Considered collectively, the mandates called for "the development and implementation of a statewide teacher assessment/evaluation system for the purpose of providing professional support for new teachers during the early year(s) of initial employment and the periodic evaluation of all Louisiana teachers for the purpose of renewable certification" (LDE, 1990). With the passage of state legislation in 1988, Louisiana became the first state

to initiate a program to assess all teachers on the job for purposes of renewable professional certification.

Requirements outlined in the Children First Act (1988) stipulated that all Louisiana teachers undergo periodic (five- year) classroom evaluations based on a "standardized process/system for the purposes of renewable state certification" (LDE, 1990). The Children First Act (1988) also contained a provision to ensure that the state teacher salary schedule would be revised and a plan developed for the Model Career Options Program (MCOP) for teachers.

The STAR was described as a "comprehensive, on-the-job teacher assessment system designed to collect information and make important decisions about the quality of effective teaching and student learning in classrooms within an interactive framework of professional development and support" (Ellett, Loup & Chauvin, 1989). The STAR was based upon an extensive review of the research literature on effective teaching (Claudet, 1990) and on an analysis and synthesis of eight large-scale teacher-performance instruments that are currently being used in a variety of statewide efforts to make decisions about beginning teacher certification, annual evaluation, career ladders, and skills needing improvement (Ellett, Garland & Logan, 1987; Logan, Garland & Ellett, 1989). While the STAR underwent slight revisions, it was basically organized by four major performance domains: I) Preparation, Planning and Evaluation;

II) Classroom and Behavior Management; III) Learning Environment; and IV) Enhancement of Learning. Each of these domains fits within a criterion of Teaching and Learning Components, which were further operationalized by sets of Assessment Indicators.

The STAR was piloted in the Spring of 1989, and the process was further developed and refined during the following year of the second pilot. Implementation of the LTIP was targeted for the 1990-1991 school year while LTEP was scheduled to be implemented statewide in 1990-1991. In October 1990, the LTIP was implemented as scheduled, but not all beginning teachers were included. The LTEP was implemented statewide, beginning in October of 1990.

The STAR was used in Louisiana to prepare school principals, master teachers, and state evaluators to conduct classroom-based assessments. While the STAR was a well-developed, solidly research- based, and psychometrically sound assessment instrument, it encountered many political hurdles. The STAR was heavily scrutinized and became the focus of severe criticism by teachers unions. As a result, it lost much of its support in both the educational and political communities. In 1991, under the Governor of Louisiana, Buddy Roemer, the Legislature suspended the assessment program. The Louisiana Department of Education was then given three years in which to revamp, pilot, and implement a new statewide teacher assessment program. Since

the STAR, Louisiana has re-developed its statewide teacher assessment system under the passage of Act 1 of the 1994 Third Extraordinary Session of the Louisiana Legislature. The current Louisiana Teacher Assessment Program is a uniform statewide program of assessment for new teachers entering service for the first time in a Louisiana Public School System and was described in further detail in Chapter One of this report.

Currently there is an increased demand for the identification of competent teachers within school systems across the nation. This demand, coupled with the availability of research and assessment instruments led to the development of large-scale teacher assessment systems which were legislatively enacted is such states as Arkansas, Connecticut, Florida, Georgia, Kentucky, Mississippi, Missouri, North Carolina, South Carolina, Tennessee, Texas, and Virginia (Chauvin & Ellett, 1991; Ellett 1990). In fact as many as eighteen states have utilized evaluation systems that were designed to include "on the job" assessment for purposes of teacher certification, merit pay, career ladders, and professional development (Association of Teacher Educators, 1988: Chauvin et al., 1991; United States Department of Education, 1987).

Observer Bias

In developing large-scale teacher assessment systems in which performance observations are essential, there are some

potential problems which need to be addressed. One of the more serious problems is that of bias, more specifically observer bias.

Research indicates that observers are sometimes not very objective in their use of observational schedules. When objectivity is not maintained, the data the observer collects tend to reflect the biases and characteristics of the observer rather than the true performance that the observational measures sought to measure (Borg & Gall 1989). This type of bias is appropriately termed "observer bias."

Observer bias refers to the systematic errors that are attributable to characteristics of the observer or the observational situation. In contrast to random errors, systematic errors are those which are made in a single direction, yielding scores that are consistently too high or too low (Borg & Gall, 1989). Observer bias may be the result of stereotypical idealogies, perceptions, and past experiences which will differ for each observer. This will lead to different perceptions of the situation, as well as different emphases, interpretation, and conclusions. Unfortunately, literature supports the notion that biases have a greater chance of operating when an observer is allowed to interpret, draw conclusions or make significant inferences from the behavior observed (Borg & Gall, 1989).

The use of rating scales and other measurement procedures that rely upon perceptions and attributions by observers may

yield data that are contaminated with what has been coined "observer bias" (Nelsen & William, 1983). It is important to remember that observer biases are not clearly visible to the observer. They can not always be seen because they involve cognitive processes which sometimes operate subconsciously within the minds of the observers. When these processes operate, it becomes difficult to extrapolate true observations of performance from those which are confounded by observer error.

Just as systems which utilize one rater can yield data that are contaminated by observer error, so can those which use multiple raters. Performance-evaluation rating systems often make use of multiple raters in an effort to improve the reliability of the ratings. Unless all candidates are rated by the same raters, some candidates will be at an unfair advantage or disadvantage due solely to being rated by more lenient or more stringent raters. Lenient raters observe performances and have a tendency to rate most individuals at the high end of the scale; these are errors of leniency. Stringent raters observe performances and have a tendency to rate most individuals at the low end of the scale (Oosterhof, 1989). It is obvious that these errors involve judgements that vary because different standards of comparison are employed by different raters (Cooper, 1981, Fiske, 1978).

While the most fruitful attempts to address the problem of low reliability have been to obtain ratings from multiple

raters by reducing the relative magnitude of random error, this process is not always fool proof (Raymond & Houston, 1990). The practice of using multiple raters does not eliminate the type of error that surfaces when individuals are evaluated by a multitude of different raters. Ratings of this nature contain systematic bias and random error. While systematic bias has been addressed, it is the random error component which is traditionally referred to as rater unreliability and which should be addressed in training.

Although most observational evaluation systems employ stringent observer training programs, they are not always effective in minimizing observer errors and biases. It is within human nature for individuals to hold pre-conceived notions and stereotypes which may contribute to observer error and biases. However, it is the intent of these rigorous training sessions to provide formal and structured instruction on how to conduct observations free from bias and to evaluate objectively.

Objectivity and Independence

The idea of "objectivity" may be misleading. Cobb (1984) spoke of the illusion of independence in evaluation which leads us to believe in the phenomenon of objectivity. In his discussion he touches on Scriven's (1976) examination of evaluator bias and independence. His translation of the word "independence" essentially suggests a meaning related to an

evaluator's distance from biasing influences; in actuality, there is a limit to this distance.

Cobb (1984) proposes that independence is valued for what it can do to achieve the two things which are most important to evaluation studies. First, the observer must be shielded from sources of bias. Secondly, independence from involvement in a program is believed to enhance the observer's credibility as an evaluator. A credible observer is assumed to be one who is capable of maintaining objectivity.

Such arguments concerning the notion of independence and objectivity are thought provoking. In fact, it is Scriven's portrayal of the misconceptions about objectivity which makes us question our modern-day philosophy of performance evaluation (House, 1980). Scriven describes objectivity as "something outside the mind that is verifiable through public or intersubjective agreement and that one can express or prove such things without influence from personal feelings" (House, 1980). An evaluation which can do so is said to be objective. However, in Scriven's final analysis, he finds fault with the prevalent misunderstanding of the principle by which most educational evaluations are guided (i.e. objectivity).

Regardless of Scriven's warning, evaluation texts often call for performance observations that are free of bias and that are objective. The reality remains, however, that raters have varying backgrounds and attitudes; therefore, they may

be biased in their observations and ultimately in their assessments. Impressions that an assessors form about an individual on one dimension can certainly influence their impressions of that person on another dimension. In many cases the assessor will express his or her overall impression of the competence of the person being rated. This practice would produce high positive correlations between ratings of presumably independent characteristics, an effect known as halo, hence the phenomenon, "halo effect" (Soar, Medley, & Coker, 1983).

The practice of stereotyping is likely. Although it may be an unconscious process, the impressions that an assessor forms about an entire group can alter his or her impressions about a group member. Frequently general conclusions are drawn about groups of people who share national origin, race, religion, gender, or other characteristics. When members of one group share such perceptions about members of another group, there are often large-scale consequences. Racial desegregation and sex discrimination are obvious examples of phenomena which owe their existence and maintenance to shared social perceptions developed from stereotyping.

Similar to stereotyping, but perhaps the most detrimental, are perception differences. These occur when viewpoints and past experiences of an assessor affect how he or she interprets behavior. Related to past experiences are knowledge and content level. When an assessor does not have

enough knowledge to make an informed judgement or decision, he or she may compensate by giving scores that are systematically higher or lower.

The importance of investigating rater and observer errors lies with understanding their impact upon performance ratings. There are numerous threats to the validity of scores based on ratings which are obtained as a result of observational rating errors. The most obvious danger lies in the meaning of the results. If ratings are obtained as a result of observer rater errors, such as the error of leniency or the halo effect, the ratings would not reflect the individuals' true performance. It is critical that an awareness exist of potential rater errors, such as those which may affect the utility and generalizability of performance observation ratings.

A recent review of evaluation literature (Martell, 1991; Dukes and Victoria, 19891; Pulako, 1989; Waldman and Avolio; 1991, Cronin and Capie, 1986) shows that the differences among raters and observers are often attributable to differences in gender, ethnicity, and even the role that the rater assumes in the assessment process. Differences attributable to gender, ethnicity and role perceptions may be the reason for existing stereotypical idealogies, differences in perceptions, attitudes, and opinions which may lead to observer bias. While other variables such as the observer's instructional level, teaching experience, and highest degree

attained have been considered, they are not nearly as serious as the gender, ethnic, and role-perception biases that the assessor may assume within the assessment process.

Gender Bias

Gender bias is said to occur when the sex (male or female) of the rater and/or ratee interferes with the rater's ability to conduct objective observations. Gender bias occur when the observer/rater incorporates his or her attitudes and perceptions about a particular sex into observation, projecting them into the ratings that they give to the individual being observed.

An extensive review of the literature addressing gender bias reveals that it occurs within many disciplines and environments. In classroom teaching situations, performance assessment is essential. However, there has been little empirical research documenting gender bias within the classroom. In a review of existing literature between 1932 and 1979, Feldman (1983) found that females were consistently rated higher than males. Female subjects also rated performance consistently higher than did male subjects (at least on some items) in studies by Basow and Distenfeld (1985); Basow and Howe (1987); Bennett (1982); and Harris (1976). However, the literature also showed evidence of approximately equal ratings for males and females in studies by Basow and Distenfield (1985), Basow and Howe (1987), and Bennett (1982). Evidence appears to be mixed on the

relationship between the subject's gender and the evaluation outcome.

In a related work, Dukes and Victoria (1989) studied the effects of gender, status, and effective teaching on the evaluation of college instruction and reported some evidence of gender bias. In this study, male subjects rated female professors higher than female subjects. In other gender research, such as that of Levenson, Bufford, Bonnoe, and Davis (1975) and Tieman and Rankin-Ullock (1985), male subjects also rated performance higher than female subjects. In all of these studies, it was found that the sex of the rater and ratee interfered with the objectivity and credibility of the rater, as well as the validity of the ratings.

A study by Goldberg in 1968, however had different results. Goldberg's (1968) study was ideally suited to the examination of teaching evaluations. It also uncovered a bias in which subjects continually favored a male over a female. Subsequent research showed that males were evaluated higher than females for the same performance and that the status of the person being evaluated altered the situation.

Gender bias not only occurs within educational systems and classroom environments, but also within industrial and non-industrial work settings. Within these settings, gender bias has been shown to affect both job and personnel evaluations as well as employment decisions.

As early as the 1970's, documentation regarding sex bias in job evaluations occurred. In Bass and Barrett (1972) and Guion (1965) the problem of subjectivity of ratings, which often played a role in the selection and promotion of personnel, was addressed. In their work, the problems and errors in ratings that affect subsequent performance evaluations were shown and were thoroughly documented.

Following the lead of Bass and his colleagues, Schmitt and Hill (1977) also studied sex as a determinant of ratings and its effect as an agent of bias. Schmitt and Hill commented that the presence of high interrater reliability did not preclude bias in the rating and that this was particularly true in the case of bias associated with sex. Schmitt and Hill found the characteristic of sex to be "reliably identified and associated with relatively strong and widely shared cultural stereotypes" (p. 261). In a series of studies employing "in-basket" techniques, Rosen and Jerdee (1973, 1974a, 1974b) demonstrated that cultural stereotypes associated with appropriate sex roles affected a variety of personnel decisions. Specifically, they found that male administrators tended to discriminate against female employees in personnel decisions involving promotion, development, and supervision. They also found that lowest acceptance rates and poorest evaluations of candidates for managerial positions went to female applicants.

Later studies addressed gender bias within employment settings more closely. In Cooper (1985), sex bias in job evaluation was studied extensively Cooper framed his study within the controversial theory of "comparable worth," which advocates that jobs of similar "worth" should be paid similarly. Within this theory, job evaluation is accorded a new role, serving as a "measure of worth and as a predictor of nondiscriminatory pay" (Remick, 1984). Unfortunately, the call for job evaluation studies such as these has undergone heavy criticism because of the possible presence of sex bias (Blumrosen, 1979; Treiman and Hartmann, 1981). Proponents of the Comparable Worth Theory have argued that the use of weights derived through the judgmental method of rating are biased against female sex-typed jobs (Blumrosen, 1979, Remick, 1981). Blumrosen's (1979) and Remick's (1981) results proved that the weights did vary as a function of the characteristics of the judges and did not vary as a function of the sex of the rater.

In a follow-up study of Cooper (1985), Arvey (1986) focused on the issues pertaining to possible sex bias in job-evaluation procedures. In this study, attention was given to possible sex bias in job analysis procedures, choice, and weighing of factors. Arvey contended that many of the procedures involved in job evaluation were inherently subjective and therefore suspect as biased and discriminatory regarding jobs held predominately by females. In support of

Arvey, Treiman and Hartmann (1981) suggested that sex stereotypes could possibly influence the nature of job-evaluation procedures and outcomes to the detriment of females.

One of the most common hypotheses encountered within the Comparable Worth literature is that female jobs are given lower evaluations than male jobs even when they are of similar value. Remick (1984) and Treiman and Hartmann (1981) hypothesize that because of stereotypes and sex bias, jobs which are populated predominantly by females are indeed systematically undervalued on job-evaluation instruments. Surprisingly, there are very few core empirical investigations of this hypothesis. The literature finds that most of the evidence cited in support of this hypothesis is studies showing lower evaluations for females in interview and performance-appraisal contexts (Arvey, 1979). The data that supports the notion that differential evaluations for male and females occur as a function of the job are used only as "direct evidence that differential evaluations of jobs occur as a function of the sex-typing of the job" (Arvey, 1986).

While Arvey was unsure if systematic sex biasing that operates "against" female-dominated jobs in job evaluation contexts actually occurred, Mount and Ellis (1987) were not. A related experimental study investigating the effects of knowledge of current pay levels and perceived job gender on

job-evaluations by Mount and Ellis (1987) proved to be quite insightful. With a sample of 53 job evaluators in professional and scientific positions at the University of Iowa, who had previously undergone 20 hours of job evaluation training and had participated in over 100 hours of job evaluations, the researchers sought to test their hypothesis. The hypothesis stated that jobs with high (manipulated) pay levels and appropriate gender would receive higher evaluations than jobs with low (manipulated) pay levels and appropriate gender. Unlike the findings of other researchers, Mount and Ellis found a marginally significant ($P<.08$) main effect for job gender, which, contrary to expectations, indicated a tendency toward pro-female bias on the part of the evaluators.

A more recent, meta-analytic review, which investigated the effects of type of sex discrimination and bias in the workplace revealed that with limited information regarding applicant competency, female applicants were evaluated more negatively than were male applicants (Gordon and Owens, 1988). It also indicated that in non-traditional work settings, women are generally hired less frequently and their work performance judged less favorably than that of men. In fact, across 19 studies and 1842 subjects, male applicants were preferred over identically qualified female applicants (Martell, 1992).

Issues of gender bias were also discussed within the counseling and psychology literature. It has been suggested that clients are subjected to gender bias in individual and cross-cultural counseling, as well as in the practice of psychotherapy. Research on counseling and psychotherapy in the United States has a brief history, yet the negative impact of sexism on counseling and psychotherapy with female clients has been described by many researchers(A-Issa, 1980; Chesler, 1972; Collier, 1982; Rice and Rice, 1973). Unfortunately the problems of sex discrimination and sexist practices within psychotherapy and counseling have not been resolved (Parloff, Waskow, and Wolfe, 1978; Sherman, 1980).

In a review of the earliest and most prominent research on sex bias in psychotherapy, Abramowitz and Dokecki (1977) suggested that the evidence of sex bias was more likely to be found in "archival data" rather than in "clinical analogue studies." These researchers proposed abandoning any experimental analogue approaches in favor of naturalistic studies of client-counselor interactions. In their discussion of instruments designed to measure attitudes and behaviors on sexism, such as the Attitude Towards Women Scale (Spence, Helmereich, and Strap, 1973, and the Bem Sex Role Inventory, Bem, 1974), Abranowitz and Dokecki believed measures such as these to be highly transparent and likely to elicit social desirability responding.

Buczek (1981) found evidence of sex-role stereotyping by therapists on an incidental memory task, a task that appeared "less reactive" to social desirability responses. In that study, 87 internship-level psychologists viewed an audiotape simulation of an intake interview of either a female or male client. Later the psychologists were asked to write down the information they recalled. In the study, psychologists were also presented with a "recognition task" and a "free response task" of generating questions to ask the client, to determine whether the counselor's information-gathering behavior reflected "traditional" sex-role stereotyping. Buczek found that the counselor's attention and "data gathering behavior" was possibly influenced by traditional sex-role stereotyping. It seemed that the counselors remembered more information about male clients and that male counselors asked female clients more questions about domestic and social concerns. This study also demonstrated that female counselors retained "significantly more client information" than their male colleagues. In addition to Buczek's work, results of memory studies with non-counselor populations also indicate the sensitivity of various memory tasks to sex-role stereotyping (Halpern, 1982; Park and Rothbart, 1982). It appears that tasks involving memory are promising measures of sex-role stereotyping or sex bias in counselor, as well as non-counselor, populations.

Outside of these areas, research has been conducted on the effects of cognitive appraisal schemes (Robbins and DeNisi, 1993) and even investigations into what role gender plays in the evaluative judgement of convention program proposals (Cooper, 1985). In Robbins and DeNisi's research, they sought to identify moderators which operate to influence sex bias in performance evaluations, as they believed sex bias to be situationally determined. Robbins and DeNisi advocated that the research on sex bias in performance appraisal needed to be integrated with the recent "cognitive" approach to understand further the evaluation process. The cognitive approach emphasizes "inaccuracy (in evaluation) as being a result of processing limitations in human nature" (Robbins and DeNisi, 1993). Prior studies have also suggested that gender may play an important role in the cognitive processing of information. Beauvais and Spence (1987) cite gender-based categorization strategies as playing a role in information processing. Hastie (1981) actually describes the relationship as gender schema which serve as "the road map for both original processing as well as the retrieval of complex information" (Freedman and Phillips, 1988, p.236). Freedman and Phillips (1988) go further by suggesting that it may be stereotypes which actually provide the basis for these gender schema.

In Cooper's (1985) study evaluating the effects of familiarity, gender, and institutional prestige on evaluative

judgements of convention program proposals, the impact of gender was measured. Cooper's analyses revealed no significant effects due to reviewer gender but did reveal severe effects due to author gender. Papers submitted by females authors were evaluated as contributing less, tending to have lower-quality analyses, and having less valid links between results and conclusions. It was also found that papers submitted by females were evaluated as having lower-quality discussions than papers submitted by males and also received lower acceptability recommendations than papers submitted by males.

Given all of these possible environments, there remains a variety of instances where gender bias is suspected and should be investigated. It is essential that we look within our educational-assessment systems for gender bias in ratings or any other form of bias.

Ethnicity Bias

Ethnicity bias occurs when the race (Caucasian, Native American, African American, etc.) of the rater/ratee interferes with the rater's ability to conduct objective observations. The rater bias is reflected in ratings and is directly attributable to the ethnicity of one or more of the individuals involved. It is probable that the rater incorporates his or her own attitudes, perceptions, and ethnic stereotypes about a particular race into observation and projects them into the ratings that they give to the

individual. Unfortunately, these ratings are not valid measures of the individuals performance and are unwarranted.

Just as gender bias occurs within varied environments outside of the realm of education, so too does ethnic bias. Ethnic bias, or racial bias as it is more commonly termed, affects many individuals' employment opportunities. The people who are most affected are those who are minorities. As early as 1974, the possibility that race served as a determinant of ratings by potential employers was studied. In Hammer et al (1974) the way that the sex and race of the "rater" and the sex and race of the "ratee" influenced assessments of ratee performance in simulated work-sampling tasks was examined. In this study, 36 undergraduates assumed the role of a manager and rated all eight combinations of male-female and black-white performers Results indicated that sex-race stereotypes did influence assessments of behavior on a work sampling task, even when objective measures were defined. In Bigoness (1976) the effects of ratee race as well as sex on rater evaluation when objective performance standards were previously established were studied. In Bigoness' study, 60 White male undergraduates in a personnel management course were assigned the role of grocery store manager and viewed a film depicting the performance of eight stockroom employees representing four sex/race combinations. Results from the study indicated that subjects could clearly distinguish between high and low performers, yet low-

performing Blacks were rated significantly higher than low-performing Whites. There was, however, no significant difference found between the subjects' ratings of high-performing Blacks and Whites.

There is considerable evidence that raters evaluate the job performance of blacks less favorably than the job performance of whites, especially when the raters are themselves white (Kraiger and Ford, 1985). There is also similar evidence that black managers experience "restricted advancement opportunities" (Alderfer, Alderfer, Tucker and Tucker, 1980; Irons & Moore, 1985; Nixon, 1985a) and report extensive "dissatisfaction and frustration" with their careers (Fernandez, 1985; Jones, 1986). Illgen and Youtz (1986) continue to suggest that minorities, especially Blacks, may experience treatment discrimination in a number of respects and that such "unfavorable experiences can have dysfunctional consequences for their career success."

In Pulakos's 1989 study examining the race and sex effects on job (military) performance ratings, racial bias was found. In this study the effects of rater source, rater and ratee race, sex, and job type were investigated on ratings collected for 8,642 first-term Army enlisted personnel. In this study, ratings were made on ten "behaviorally based" dimensions that had been developed for first-term soldiers. Results revealed a significant main effect and interaction effect for sex as well as race. It

was found that race, as expected, was a determinant in the type of ratings that were received. While Pulakos' and his colleagues consistently found significant race effects, these effects explained extremely small amounts of variance in ratings.

Waldman and Avolio's 1991 study of race effects within performance evaluations also produced somewhat similar results. Waldman and Avolio examined the effects of ratee and rater race (Black or White) on performance evaluation ratings of 21,000 individuals employed in ten occupational categories. Using hierarchial regression analyses, the researchers found a significant main effect attributable to ratee race, although the magnitude of this effect was varied across occupational types. Waldman and Avolio found no evidence of a same race (i.e., rater-ratee) interaction effect as Pulakos (1989) did. They found that after individual differences in ability and length of experience were controlled, the race of the rater and the ratee accounted for little variance in performance evaluations. Waldman and Avolio recommended that future researchers examine the "qualitative experiences" of White and Black employees to determine what might account for differences in the group's performance.

Another recent study which attempted to determine the effects of race on organizational experiences, job performance evaluations, and career outcomes also generated

significant findings in race effects. Greenhaus, Parasuraman & Wormley (1990) examined the relationships among race, organizational experiences, and career outcomes for 455 White and 373 Black managers. Each subject's supervisor also participated by evaluating job performance and career outcomes. Compared to Whites, Blacks were reported to feel less accepted in their organizations and also perceived themselves as having less "discretion" on their jobs. In comparison to Whites, Blacks also received lower ratings from their supervisors on their performance and promotability, were more apt to have reached "career plateaus," and experienced lower levels of "career satisfaction."

An earlier study (Lawrence et al 1987), addressing supervisory ratings of employees, attempted to identify factors related to perceptions of performance. In the study, were identified related to Black/White and Female/Male employees' perceptions of the accuracy of ratings made using a "subjective" rating system. The sample included 103 White females, 24 Black females, 98 White males and 9 Black males. The researcher's analysis revealed factors dealing with the relevance of the appraisal instrument and confidence in their supervisors' qualifications to "accurately" rate the employee's performance, as well as matters related to appraisal outcomes" (i.e., rewards and career advancement). The researchers regressed the measure of perceived fairness and accuracy onto three individual factors and for race/sex

and demographic variables. Results indicated that race was indeed related to perceived fairness and accuracy.

Following the work of Lawrence is an extension of the research on supervisory job evaluations. In Greenhaus and Parasuraman (1993), the impact of the manager's gender and race on job performance attributions made by supervisors was examined. Greenhaus and Parasuraman surveyed 748 managers (211 Black women, 124 Black men, 212 White women, and 193 White men) and their supervisors. The performance of Black managers was perceived to predict for them less "favorable career advancement prospects than White managers." The researchers concluded that the effect of race on "career advancement prospects" was thought to be "direct," as exhibited through performance ratings and attributions of ability.

Related to this research on supervisory performance ratings is a recent extensive meta-analysis of eight separate studies. In Martocchio and Whitener (1992) the results of the meta-analysis of the studies, involving ten "independent" samples, indicated that Whites performed higher than non-Whites on supervisory ratings, but not on objective results. In their discussions of the validity of performance measures, Martocchio and Whitener suggest that minorities may be receiving lower performance appraisals due to prejudice or bias on the part of their supervisors. Several researchers

have since explored the extent to which performance measures, particularly supervisory ratings, might be biased.

Kraiger and& Ford (1990) also examined the effects of a ratee's race on the relations between supervisory ratings and more objective criteria such as "job knowledge" and "work performance." Like Martocchio and Whitener, Kraiger and Ford performed a meta-analysis of 12 studies published between the years of 1960 and 1988. The results of their meta analysis also supported theories of ethnic bias. In fact, supervisory ratings were "more highly related to work-performance measures and to a lesser extent to job-knowledge measures for Black than for White ratees." Kraiger and Ford suggested that these differences could be accounted for by "inter group theory" and "positivity bias theory," which are theories for which higher ratings for same-group ratees are given. Previously in 1985, Kraiger and Ford had performed a meta-analysis in which Black and White raters evaluated Black and White ratees. Even then, they found that raters tended to evaluate ratees of their own race higher than ratees of the opposite race. Kraiger and Ford noted that since supervisors are predominantly White, this result clearly suggests that minority ratees may be receiving lower evaluations

In sum, research addressing issues of ethnicity and bias within assessment and evaluation is limited. However, most of the literature suggests that minority teachers face negative effects of evaluations of any kind. Minority educators are

extremely concerned with fair, accurate, and productive evaluations (Peterson, Dehyle and Watkins, 1988).

Role Perception Bias

Traditionally, the principal has been seen as the instructional leader in schools (e.g., Levine & Lezolte, 1990). Moreover, when principals are surveyed, they list instructional leadership as their most important role (Lane 1990). The other demands on their time, such as conducting personnel evaluations, usually relegate active leadership of the instructional program to a minor role. When this happens, many principals experience role conflict. While they still regard the role of instructional leader to be their major responsibility, they realize that to exert leadership they must deal with supervision and staff development as well as personnel evaluation. They often are not prepared to serve as primary evaluators. Principals many times experience role-perception bias, which interferes with their ability to conduct proper and effective teacher evaluations (Medley and Cocker, 1987).

When applying the phenomenon of role perception bias to teacher assessment systems, it is intended to mean the way a person comes to view his identity as an assessor and its effect on his/her ability to carry out the task of assessment. Role-perception bias occurs when the function that individuals serve and the role that they assume within

the observational schema affects their ability to observe objectively.

Research suggests that the role a person serves within an observational-assessment process is critical and must be clearly understood (Ellett and Capie, 1985; Acheson, Smith and Stuart, 1986; Garland, 1989). Literature addressing the many roles and conflicts of persons who conduct observations to evaluate is abundant (Acheson, Smith & Stuart, 1986, Price, 1989; Collins, 1990). One of the most obvious conflicts is felt by the building principal who is expected to be a hard-nosed evaluator of teachers as well as a kind, sympathetic, and helpful supervisor of instruction (Acheson, Smith and Stuart, 1986). Many times principals can not balance these roles. The pressure to balance the roles of supervisor and evaluator can be overwhelming. This tension and the problems of limited time and insufficient experience to supervise and evaluate teachers lead many researchers to conclude that it would be best to separate the role of clinical supervisor or helpful colleague from the role of evaluator (Price, 1989).

In a critiques of current evaluation practices, Collins (1990) discusses the dilemmas that arise when principals assume both responsibilities of evaluation and supervision roles. Collins describes one of the most persistent problems in supervision as the dilemma between (a) evaluating a teacher in order to make decisions about retention,

promotion, and tenure and (b) working with the teacher as a friendly critic or colleague to help develop skills the teacher wants to use and to expand the repertoire of strategies that can be employed. Collins later summarizes that the task of both supervising and evaluating teachers demands an especially delicate balance. He also believes that effective supervision requires a high level of trust, yet teachers often regard any evaluation that is less than laudatory as an "attack" on their character. To evaluate them as teachers is to evaluate them as persons and in Acheson, Smith and Stuart's (1986) words, "it is difficult to trust someone who is (in your mind) slandering and defaming your heart and soul."(p.6).

Additional literature suggests that the principal (or immediate supervisor) may experience role conflict in trying to serve as the instructional leader and as an administrative decision maker (Duckett, 1985; Stanley and Popham, 1988). It is very difficult for a principal to make employment decisions based on the outcome of formal evaluation, when he or she wishes to see that teacher improve instruction. To make administrative decisions is cumbersome when the principal has a stake in how a teacher performs during evaluation. After all, it is the principal who chose the new teacher, and to have the teacher fail would not be a good reflection of the principal's judgement.

Previous research addressing on-campus (principals) and off-campus (university faculty members) evaluators has also become an issue of concern. Findings indicate that on-campus evaluators/principals tend to award higher scores than off-campus evaluators or those who do not directly serve as the teacher's immediate supervisor (Cronin and Capie, 1986; Ellett and Capie, 1985; Ellett, Teddlie and Niak, 1991; Kelly, 1985; Rose and Huynh, 1984; Wise et al., 1984). As has been alluded to earlier, principals want their new teachers to succeed and continue employment. Principals show allegiance to their young teachers and may not be able to evaluate objectively. In situations where principals could and "must" are to observe and rate performances of new teachers, objectivity is even harder to maintain.

Research indicates that most observers, including principals, are sometimes not very objective in their use of observational schedules. When objectivity is not maintained, the data the observer collects tend to reflect the biases and characteristics of the observer rather than the true performance that the observational measures sought to measure (Borg and Gall 1989). In Medley and Coker's (1987) article, they question the principal's objectivity and seek to determine the validity of the principal's judgments of teacher effectiveness. Medley and Coker admit that the question of whether or not a principal's judgements are valid is a natural and important one, but one that is rarely

asked. In fact, they feel that the validity of any evaluators' judgements are generally taken for granted. However, in the limited number of studies that have been conducted addressing the validity of the principal's judgments (or of ratings based on them) that have been reported in the literature, there have been consistently negative findings (Medley and Coker, 1987).

The results of Medley & Coker's study has supported the literature, as they found "low accuracy of the...principal's judgments of the performance of the teachers he or she supervises" (p.245). Medley and Coker also addressed whether or not anything could be done to make principal's ratings more responsive to teacher effectiveness. However, their response that "the strength of the overall impression that a principal forms of the effectiveness of any teacher is so strong that it is doubtful that any amount of training can overcome it" (p. 140) does not seem promising. Within the same year, Peterson (1987b) reported on correlations between administrator ratings and various other teacher data sources. Peterson found a correlation of only 0.01 between administrator ratings and those of students, 0.09 for administrators and parents, -0.12 for administrators and test scores, and -0.11 between administrators and documentation of professional activity. These findings corroborated those of over 11 earlier studies reviewed by Medley and Coker (1987).

It is believed that administrators ratings of teachers are also inaccurate because of sociological reasons (Peterson et al, 1988). While principals have access to considerable data about teachers, they are also under "considerable" pressure from their faculties to get along day to day (Lortie, 1975). Peterson believes commendations and criticisms are used in a school workplace in exchange for cooperation, support, reward, control, and even punishment. Peterson's philosophy is that although it is most often unintentional, administrators' ratings are compromised in accuracy and specificity. Principal evaluations are believed to have only continued because of precedent and a lack of any demonstrated alternative.

While present traditional evaluation is dominated by administrator ratings using checklists (Lewis, 1982), this assessment procedure results in "doubtful validity and reliability for both technical and sociological reasons" (Darling-Hammond et al., 1983). Researchers have consistently found a pattern of inaccuracy in principal ratings of teachers for over 30 years of research in the field (Medley and Cocker, 1987), and this fact is unsettling.

Acheson's (1986) writings also point to other reasons why principals have difficulty using observational methods to evaluate teachers. He reports that there may be a lack of knowledge of the range of observational techniques, a lack of understanding of instructional strategies and a lack of

interpersonal skills used by principals as observers in providing feedback to teachers.

Clearly the persistent problem of performing in a supervisory and evaluatory capacity remains a dilemma for principals and administrators alike. This dilemma is magnified when principals are called upon to conduct performance observations and must be resolved.

As the literature reveals, issues of gender, ethnic, and roles biases are prevalent within many disciplines. While there have been a number of studies conducted over the last 25 years, they only provide small pieces to a much larger puzzle. There is a real need for a study that systematically incorporates gender, ethnicity, and role effects and which uses more than one methodological approach (e.g. causal comparative and experimental) and more than one mode of inquiry (e.g. qualitative and quantitative). This is the intent of this research.

CHAPTER 3: METHODOLOGY

This chapter begins with an overview of the assessment instruments and data used in the study. The data come from two studies. Study I was causal comparative in nature, using the pilot assessment data from the LTAI. Study II was an experimental study using effectiveness ratings. After a discussion of the data, the sampling procedure is described. A section delineating the various variables and measures used to answer the research questions for this study follows. The last section of this chapter is a discussion of the design of the study and the methods to be employed in analyzing data.

Study I

Instruments and Data

The Louisiana Teacher Assessment Instruments (LTAI) were administered during the 1993-94 Pilot Test of the Louisiana Teacher Assessment Program for Interns. The LTAI consists of a preobservation conference interview, a classroom observation, and a postobservation conference. The preobservation interview and the classroom observation used instruments and procedures designed to collect data related to the Louisiana Components of Effective Teaching (LCET) (See Appendix B).

The LCET are hierarchal, with three levels of teaching skills and knowledge that form the assessment criteria. The LCET have been developed from the professional knowledge base

on teaching and "craft knowledge" acquired by experienced educators (LDE, 1994). A panel of educators (Panel 1) reviewed the professional knowledge base on teaching by examining research-based teacher assessment and evaluation documents from eight states. In addition, other experts on personnel evaluation were consulted. Recommendations were also received from out-of-state and in-state review teams. Panel 1 also used the position paper of the Teacher Evaluation Advisory Commission to develop the criteria (LDE,1994).

A content validation study was done on the LCET by educational consultants employed by the Louisiana Department of Education. The results of the study suggested that the attributes, components, and domains which form the basis of the LCET were valid indicators of effective teaching (Doescher and Brooks, 1993).

The data for the study were the assessment ratings collected during the 1993-1994 Pilot Test of the Louisiana Teacher Assessment Program for Interns (LTAPI) and were available at the rater (assessor) level. For each assessee, there were performance observation ratings from each of three assessors across all 27 attributes, 8 components, and 4 domains. The LTAI used a three-point rating scale designed to allow formative feedback to the intern teacher. The three-point rating scale was used in all instruments to determine ratings on all components and attributes during each

individual assessment visit. The three points were defined as 1) needs improvement, 2) proficient, and 3) commendable (LDE, 1994).

Assessment data were merged with assessor and assessee demographic records to produce one comprehensive data set. Data from the LTAI were available for 404 of the 430 interns and for all of the 721 assessors. The unit of analysis for this study was the assessee and the assessor.

The process by which ratings were attained is multi-dimensional. All attributes were rated using a three-point scale. Individual assessors then combined these attribute ratings into component ratings. In formulating the component ratings, the assessor analyzed the pattern of attribute ratings for that component and "determined the component ratings most representative of the pattern, taking all practices and behaviors into account" (LDE, 1993).

For purposes of this study, a slightly modified version of the component scoring scale was utilized. To protect the assessors' original scoring rationales, a component score was constructed by summing the attribute ratings.

Sample

The target population for the 1993-94 Pilot Study included all teachers teaching in Louisiana's public schools. In selecting the sample for the Pilot Test, **Local Educational Agencies (LEAs)** were identified based on the projected number of interns (Bulletin 1472, 1991-1992 Annual Financial and

Statistical Report), representation from each of the eight Regional Service Centers, geographic proximity for assessor training and intern orientation, gender, and ethnicity. If an LEA agreed to participate, all interns in the LEA were included in the Pilot Test. Originally, 13 LEAs agreed to participate. However, in these districts the actual number of interns was lower than projected. To augment the sample, four additional LEAs were recruited to participate; therefore, all the interns from these four LEAs were included (LDE, 1993). Furthermore, interns were sampled from three other LEAs. The final sample consisted of 430 interns from 20 LEAs participating in the Pilot Test. Table 1 shows the number of participating interns in each LEA.

Assessors were also recruited from within the 20 LEAs and were required to undergo extensive six-day training to become knowledgeable and reliable assessors. Assessors were of varied backgrounds. They were principals, assistant principals, experienced teachers, retired administrators, retired teachers, central office personnel, university faculty and state department officials. Almost all assessors completed assessor training successfully. Those assessors who were not directly employed by the LEA and who were assigned to assess intern teachers were paid a small stipend.

Existing data represent assessment ratings collected during the 1993-1994 Pilot Test. There were 404 interns and 721 assessors. Of the assessors reporting their work areas,

Table 1

Intern Participants by Regional Service Centers (RSC)

RSC	Number of Teachers
RSC I	
Jefferson	13
Orleans	5
St. Bernard	16
St. John	6
RSC II	
Ascension	20
East Baton Rouge	16
Tangipanoa	15
RSC III	
Lafourche	18
RSC IV	
Evangeline	7
Iberia	32
Lafayette	33
RSC V	
Beauregard	21
Jefferson Davis	24
RSC VI	
Rapides	23
RSC VII	
Bossier	34
Caddo	60
RSC VIII	
Franklin	8
Monroe	15
Morehouse	14
Ouachita	18
TOTAL	404

were principals or assistant principals, 293 were experienced teachers, and 144 were external assessors. Of the assessees who reported their gender, 333 (82%) were female and 57 (14%) are male (Table 2). The percentages of females and males in the sample of assessees (interns) approximate the respective percentages of Louisiana's public school teacher population in which approximately 82% of the teachers are female and 18% male (LDE, 1994). Of the assessors who reported their gender, 468 (65%) were female and 195 (27%) were male (Table 3). Of those assessees who reported their ethnicity, 341 (84%) were Caucasian and 47 (12%) were African American (Table 2). Of those assessors who reported their race, 497 (69%) were Caucasian and 158 (22%) were African American (Table 3).The majority of assessees held either a BA or BS degree, although 5% held graduate degrees. About 53% were teaching in elementary schools, 16% in middle schools, and 19% in secondary schools (Table 2). Approximately 60% of the assessors were teaching in either lower or upper elementary schools; 17% secondary subject areas; and 14% in special education (Table 3). The majority of assessors held a master's degree plus 30 graduate hours (52%), with 30% having their master's degree only and 8% who had a Ph.D. Most of the assessors were working in elementary school settings (Table 3).

Table 2

Demographic Data for Interns in Study I

	Number of Teachers	Percent
Gender		
Male	57	14%
Female	333	82%
Ethnicity		
Caucasian	341	84%
African American	47	12%
Educational Degree		
BA/BS	370	92%
MA/MA+30	19	5%
Ph.D.	1	2%
Teacher's Type of School		
Elementary School	215	53%
Middle /Jr.High School	65	16%
Combination	29	7%
High School	75	19%

N=404

Table 3

Demographic Data for Assessors in Study I

	Number of Teachers	Percent
Gender		
Male	195	27%
Female	468	65%
Ethnicity		
Caucasian	497	69%
African American	158	22%
Educational Degree		
BA/BS	15	2%
MA/MA+30	553	77%
Ph.D.	95	8%
Type of School		
Elementary School	297	41%
Middle /Jr.High School	104	15%
Combination	41	6%
High School	118	17%

N=721

Variables

The independent variables for Issue 1 were the gender (Male/Female) of both the assessor and assessee. The independent variables for Issue 2 were the ethnicity (African American/ Caucasian) of both the assessor and assessee. For purposes of this study, only those whose ethnicity was African American or Caucasian were used in the sample. Data regarding the ethnicity and gender of the assessee and the assessor were collected on a demographic data form prior to assessment (see Appendix C).

The independent variable for Issue 3 was the role of the assessor. An assessor had only one of three possible roles: principal, experienced teacher, or external assessor.

The dependent variables for all three issues were the assessee's performance observation ratings, which were the mean ratings of the 8 Components.

Design

The design of Study I was causal comparative. This method, employed to discover causal relationships between variables, addressed possible causes for the differences in observational performance ratings by comparing assessees and assessors for which a characteristic was present (e.g. gender=female), with similar assessee and assessors for which the characteristic of interest was absent (e.g., gender not female). Unfortunately, the causal comparative method can

only be used to explore, not confirm, casual relationships (Borg and Gall, 1989).

The design of Study I was dependent on the issue, and Study I had three issues. As will be discussed later, a 2x2x2x2x3 ANOVA design was not suitable due to the many meaningless higher-order interaction effects it produces. Issue I, gender bias, was analyzed using a 2x2 factorial design with 2 levels of assessee gender (male and female) and 2 levels of assessor gender (male and female). Issue II, ethnicity bias, was analyzed using a 2x2 factorial design with 2 levels of assessee ethnicity (African American and Caucasian) and 2 levels of assessor ethnicity (African American and Caucasian). Issue III, role bias, was analyzed using a one-way design with 3 levels of assessor role (principal, experienced teacher, external assessor).

Data Analysis

With the formation of several groups and several dependent variable measures, the appropriate statistical technique to determine whether the groups differed in performance ratings was a multivariate analysis of variance (MANOVA). A MANOVA was used to determine the effects of the independent variables for each issue on the combination of all eight dependent variables (Components). The MANOVA produced F ratios for the effects of each independent variable on the dependent variables (main effect). The MANOVA also produced two 2-factor interaction effects. The 2-factor

interaction effects were assessor/assessee ethnicity and assessor/assessee gender.

All significant multivariate tests were followed by separate univariate tests of significance for each variable. Essentially, if the MANOVA produced significant effects, corresponding ANOVAs were performed on each of the eight dependent variables.

Study II

Instruments and Data

In addition to the analysis of the assessment data from Study I for evidence of bias, an experimental study was also conducted to study the perceptions and attitudes of effective teaching behaviors and practices. Study II consisted of a detailed description of a classroom teaching scenario which was to be rated for effectiveness by assessors. (see Appendix D). After the assessors read the scenario, they were asked to rate the fictitious teacher's perceived level of effectiveness on a 4-point Likert-type scale, with one indicating "not effective" and four indicating "highly effective." The instrument also contained open-ended questions concerning the teacher's level of effectiveness and a section for the "assessor" to report demographic information.

This teaching scenario, an experimental stimulus, was developed by the researcher and three Louisiana Department of Education staff members, each with over 20 years of classroom

teaching experience. They are proficient in the LCET. The scenario incorporated several direct and indirect examples of both effective and ineffective teaching behaviors and attitudes. The LCET were used as a guideline in developing the scenario to portray a "moderately effective" teacher.

Experimental manipulation consisted of presenting a scenario to the respondent. Each version had a description of an an elementary science teacher. The gender and ethnicity of the portrayed teacher was systematically manipulated, while all other information was held the same across experimental conditions. For example, each respondent received one of the four scenarios depicting either a White female teacher, a White male teacher, a Black female teacher, or a Black male teacher. Gender and ethnicity of the respondents receiving each of the four versions of the scenario were identified, by assigning the material several numbers before mailing.

Sample

The sample for this study was those assessors who participated in Study I and whose gender, ethnicity, and role were identified from the previous data. Permission to have their school personnel participate in the experimental study was requested of the initial 20 LEA superintendents.

The sampling scheme can be found in Table 4. As that table indicates, approximately 25 assessors were sampled in

each of the 16 types of teacher/assessor combinations. This sampling scheme was used to ensure that approximately 15 respondents returned their ratings in each of the 16 cells. Each respondent was sent only one instrument to rate.

Due to the unequal numbers of available assessors with certain gender and race characteristics (e.g., Black males and Black females), in some groups of assessors, the ratio of selected members was larger than in others. Since each of the groups of sampled assessors differed in number, even with oversampling, care was taken to keep size discrepancies between the four groups to a minimum.

Of the 20 LEAs, 14 (70%) agreed to participate in the study. Of the 471 assessors, close to 73% (n=348) were sampled. Approximately 47% (n=162) responded to the mailed instrument. Follow-up letters to those assessors from the primary sample who did not return the instrument resulted in an additional 15% (n=52) response. The follow-up letter was used only for cells in which the initial response rate was too low and there were less than eight responses. The final number of returned instruments was 214, or 65%. The response pattern can be found in Table 5.

Of the 214 assessors who returned the effectiveness questionnaire, 118 (55%) were female and 96 (45%) were male (Table 6). Also, 125 (58%) assessors were White and 89 (52%) were African American.

Table 4

Study II: Cases in Experimental Conditions

	4 STIMULUS TEACHER TYPES				
ASSESSOR TYPE	Black Female	White Female	Black Male	White Male	TOTAL
Black Female	n=25	n=25	n=25	m=25	N=100
White Female	n=25	n=25	n=25	n=25	N=100
Black* Male	n=12	n=12	n=12	n=12	N=48
White Male	n=25	n=25	n=25	n=25	N=100
Total	N=87	N=87	N=87	N=87	N=348

* The number of African American Assessors in the 14 LEAs was 48.

Table 5

Study II: Returned Responses

	4 STIMULUS TEACHER TYPES				
ASSESSOR TYPE	Black Female	White Female	Black Male	White Male	TOTAL
Black Female	n=12	n=14	n=13	m=14	N=53
White Female	n=13	n=17	n=19	n=16	N=65
Black Male	n=9	n=9	n=9	n=9	N=36
White Male	n=15	n=17	n=13	n=15	N=60
Total	N=49	N=57	N=53	N=54	N=214

Table 6

Study II: Demographic Characteristics of Respondents

	Number of Teachers	Percent
Gender		
Male	96	45%
Female	118	55%
Ethnicity		
Caucasian	125	58%
African American	89	42%
Age		
26-30	2	1%
31-35	14	7%
36-40	45	21%
41-45	64	30%
46-50	49	23%
51-55	22	10%
55 and Over	13	6%
Years Teaching Experience		
0-10 Years	36	17%
11-20 Years	102	47%
21 Years (+)	112	36%

While the age of the respondents varied, more than 70% were over the age of 40, and approximately 65% had over 15 years of classroom teaching experience.

Variables

The independent variables for Issue 1 were the gender (Male/Female) of both the assessor and assessee. The independent variables for Issue 2 were the ethnicity (African American/Caucasian) of both the assessor and assessee. Ethnicity of the assessee was manipulated in the teaching scenario. For purposes of this study, only those whose ethnicity was African American or Caucasian were used in the sample.

The dependent variable for all three issues in Study II was the teacher's effectiveness rating. The rating was based on a Likert-type scale ranging from 1 to 4, with a rating of 1 labeled as "Not Effective" and a 4 as "Highly Effective". The question on the instrument was, "In your opinion, how effective is Mrs. Jones's (or Mr. Jones's) teaching?"

Design

Study II was an experimental study designed to follow and expand upon the findings of Study I. As an experimental study, Study II involved actively manipulating independent variables. In this study, both the gender and ethnicity of the teacher in the scenario were manipulated. In Study I, there were very few assessment situations that provided data

regarding Black male assessors or even assessees. Study II was designed to deal with that problem.

Data Analysis

The data were analyzed for evidence of assessor gender and ethnicity bias through the use of ANOVA. A 2X2X2X2X3 ANOVA (Assessor Gender by Assessee Gender by Assessor Race by Assessee Race) was performed on the effectiveness data.

Endnotes

At the conclusion of the assessment process, the intern's team of assessors combine the component ratings from the three visits, using a two-point rating scale (1 = Fail, 2 or above = Pass) to create a certification decision. The assumption is that those who receive a 2 or above on all their component ratings will be recommended for certification and are considered competent teachers. Since the purpose of the present study is to study gender/ethnicity and role bias in ratings, the original 3-point scale is used instead of the 2-point pass/fail rating scale.

In at least one of the groups, such as African American males, the number of assessees was to small. It is for this reason that a 2X2X2X2X3 ANOVA was not used.

CHAPTER 4: RESULTS AND DISCUSSION

The results of Study I and Study II are presented in this chapter in both tabular and narrative forms to address the research questions of interest. The research questions for both Study I and Study II are addressed in the order that they were first presented in Chapter I. All analyses were performed through the use of the SPSS Data Analysis System, Release 4.0 (SPSS, 1990).

Results for Study I

Reliability Estimates of Internal Consistency

As mentioned in chapter 3, the reliability of the assessment instrument and its items is important and should be addressed. The relationship between individual attributes and the composite component scores was investigated in order to interpret correctly any future analyses. In Tables 7 through 14, the reliability estimates of each of the eight component scales, including the item-total statistics, are presented.

All the attributes (items)and their corresponding composite values (components) appeared to be reliable. The squared multiple correlation coefficients for each of the attributes were low to moderate with coefficients ranging from .22 to .59. As shown in Tables 7 through 14, the squared multiple correlation coefficients (R^2)differed within the various components.

Table 7a

Reliability Estimates of Component One: Teacher Plans Effectively for Instruction

ATTRIBUTE	X	SD	*R^2
Specifies learner outcomes in clear, concise objectives	2.50	0.52	0.46
Includes activities that develop objectives	2.54	0.53	0.52
Identifies and plans for individual differences	2.28	0.62	0.36
Identifies materials other than standard classroom materials as needed for lesson	2.48	0.56	0.49
States methods of evaluation to measure learner outcomes	2.40	0.54	0.37

*R^2 =Squared Multiple Correlation
ALPHA= .84

Table 7b

Reliability Estimates of Component One: Teacher Plans Effectively for Instruction (For Special Education)

ATTRIBUTE	X	SD	*R
Specifies learner outcomes in clear, concise objectives	2.51	0.53	0.51
Includes activity/activities that develop objectives	2.52	0.55	0.40
Identifies materials, other than standard classroom materials, as needed for lesson	2.48	0.55	0.51
States methods of evaluation to measure learner outcomes	2.44	0.53	0.36
Develops an Individual Education Plan	2.44	0.54	0.59

*R =Squared Multiple Correlation
ALPHA=.84

Table 8

Reliability Estimates of Component Two: Teacher Maintains Environment

ATTRIBUTE	X	SD	*R
Organizes available space, materials, and/or equipment to facilitate learning	2.64	0.51	0.27
promotes a positive learning climate	2.63	0.52	0.27

*R =Squared Multiple Correlation
ALPHA= .68

Table 9

Reliability Estimates of Component Three: Teacher Maximizes Time

ATTRIBUTE	X	SD	*R^2
Manages routines and transitions in a timely manner	2.55	0.56	0.40
Manages and/or adjusts allotted time for activities planned	2.42	0.62	0.40

*R^2=Squared Multiple Correlation
ALPHA=.78

Table 10

Reliability Estimates of Component Four: Teacher Manages Learner Behavior

ATTRIBUTE	X	SD	*R^2
Establishes expectations for learner behavior	2.49	0.58	0.48
Uses monitoring techniques to facilitate learning	2.49	0.58	0.48

*R^2=Squared Multiple Correlation
ALPHA=.82

Table 11

Reliability Estimates of Component Five: Teacher Delivers Instruction

ATTRIBUTE	X	SD	*R
Uses techniques which develop lesson objectives	2.43	0.60	0.41
Sequences lesson to promote learning	2.48	0.56	0.48
Uses available teaching materials to achieve lesson objectives	2.47	0.59	0.28
Adjusts lesson when appropriate	2.30	0.64	0.36

*R =Squared Multiple Correlation
ALPHA=.79

Table 12

Reliability Estimates of Component Six: Teacher Presents Content

ATTRIBUTE	X	SD	*R
Teacher Presents Content	2.53	0.54	0.43
Presents accurate subject matter	2.59	0.52	0.48
Relates relevant examples, unexpected situations, or current events to the content	2.40	0.64	0.23
Answers questions correctly and/or directs students to additional sources	2.53	0.53	0.48

*R^2=Squared Multiple Correlation
ALPHA=.79

Table 13

Reliability Estimates of Component Seven: Teacher Provides for Student Involvement

ATTRIBUTE	X	SD	*R'
Accommodates Individual Differences	2.25	0.66	0.31
Demonstrates ability to communicate effectively with students	2.51	0.56	0.27
Stimulates and encourages higher order thinking at the appropriate developmental levels	2.27	0.64	0.27
Encourages Student Participation	2.56	0.57	0.29

*R' =Squared Multiple Correlation
ALPHA=.74

Table 14

Reliability Estimates of Component Eight: Teacher Assessees Student Progress

ATTRIBUTE	X	SD	*R
Monitors ongoing performance of students	2.47	0.55	0.43
Provides timely feedback to students regarding their progress	2.45	0.58	0.43
Uses assessment Techniques Effectively	2.47	0.54	0.47
Monitors Ongoing Performance of Students	2.44	0.52	0.47
Provides timely feedback to students regarding their progress	2.52	0.54	0.49

*R =Squared Multiple Correlation
ALPHA=.85

As seen in Table 15, measures of reliability for scales or components, as indicated by Cronbach's alpha, were moderate to high with coefficients ranging from 0.68 to 0.85. The highest alpha (0.85) pertained to component eight (Teacher Assesses Student Progress), which had an alpha of 0.85, followed by component one (Teacher Plans Effectively for Instruction), which had an alpha of 0.84 and component. The lowest alpha (0.68) pertained to component two (Teacher Maintains Environment). Given the size of the scales, these reliability indicators were fairly satisfactory.

Issue I: Gender Bias

Research Question 1. Do male and female assessors differ in their overall ratings of assessees?

Hypotheses 1a. Female assessors rate both male and female assessees higher than male assessors.

Table 16 presents the mean of the component scores of effective teaching for the four groups related to research question number one. It should be noted that since the number of items within each component was not the same, the component means were not comparable across components.

As Table 16 indicates, female assessors gave an overall scale mean rating of 8.52, and male assessors gave an overall scale mean rating of 8.43. The overall scale mean was the mean of all eight components' ratings. Male and female assessors differed significantly in their overall mean

Table 15

Overall Reliability Estimates

Component Scale	Alpha
Planning Effectively for Instruction	0.84
Planning Effectively for Instruction (Special Education)	0.84
Teacher Maintains Environment	0.68
Teacher Maximizes Time	0.78
Teacher Manager Learner Behavior	0.82
Teacher Delivers Instruction	0.79
Teacher Presents Content	0.79
Teacher Provides for Student Involvement	0.74
Teacher Assesses Student Progress	0.85

Table 16

Mean Component Ratings (Assessor/Assessee Gender)

		Assessor Gender					
		Male			Female		
Assessee Gender		**Male**	**Female**	**All**	**Male**	**Female**	**All**
Plans Effectively	**X**	**11.61**	**12.09**	**11.96**	**11.56**	**12.42**	**12.33**
	S	1.94	2.16	2.13	2.03	2.16	2.16
Maintains Environment	**X**	**4.92**	**5.33**	**5.26**	**5.06**	**5.30**	**5.27**
	S	.966	.838	.876	.952	.904	.910
Maximizes Time	**X**	**4.77**	**5.04**	**5.00**	**4.96**	**4.96**	**4.96**
	S	1.10	.954	.987	1.09	1.10	.104
Manages Behavior	**X**	**4.71**	**5.10**	**5.03**	**4.75**	**4.97**	**4.84**
	S	1.14	.984	1.03	1.05	1.08	**1.09**
Delivers Instruction	**X**	**9.12**	**9.92**	**9.78**	**9.40**	**9.93**	**9.87**
	S	1.75	1.79	1.81	1.81	1.91	1.91
Presents Content	**X**	**9.78**	**10.17**	**10.10**	**9.97**	**10.02**	**10.03**
	S	1.71	1.67	**1.69**	1.66	1.85	1.83
Student Involvement	**X**	**8.83**	**9.76**	**9.60**	**9.14**	**9.68**	**9.63**
	S	1.83	1.69	1.76	1.92	1.83	1.86
Student Progress		**11.66**	**12.38**	**12.25**	**11.60**	**12.52**	**12.42**
		1.98	2.14	2.13	2.08	2.15	2.16
Overall Scale Mean		**8.14**	**8.72**	**8.43**	**8.31**	**8.73**	**8.52**
		1.13	**1.53**	**1.17**	**1.59**	**1.62**	**1.27**

ratings of assessees, [F(8,1037)=0.646;p<.740)], although the difference was small.

The overall scale means presented in Table 16 supported hypothesis 1a, demonstrating that female assessors did rate both male and female assessors higher than male assessors. The overall scale mean rating given by female assessors to male assessees was 8.31 and to female assessees was 8.73. The overall scale mean rating given by male assessors to male assessees was 8.14 and to female assessees was 8.72.

However, when the mean ratings for each of the eight components were analyzed individually, the results were somewhat different. Male assessors gave female assessees higher ratings on 5 of the 8 components (Teacher Maintains Environment, Teacher Maximizes Time, Teacher Manages Learner Behavior, Teacher Presents Content, and Teacher Provides for Student Involvement), whereas female assessors gave female assessees higher ratings on only 3 of the 8 components (Teacher Plans Effectively for Instruction, Teacher Delivers Instruction, and Teacher Assesses Student Progress). It is apparent that female assessees are rated highest when assessed by male assessors. This finding did not entirely support hypothesis 1a.

Despite these trends, a Multivariate analysis of variance (MANOVA) did not reveal a significant main effect of assessor gender [F(8.1037)=0.646;p<.740] (Table 17). Hence, Hypothesis 1a should be considered unsupported.

Research Question 2: Is there a difference in the average ratings of male and female assessees?

Hypothesis 2a: There is a statistically significant difference in the mean ratings received by male and female assessees.

As Table 16 indicates, there was a difference in the average ratings of male and female assessees although the difference was small. On the average, the component ratings were consistently higher for female assessees, as compared to male assessees. Female assessees had an overall scale mean rating of 8.73, whereas male assessees had an overall scale mean of 8.23. When considering the eight mean component ratings individually, female assessees scored higher than male assessees across every component. This trend was present regardless of the assessor gender. The difference between male assessee and female assessee mean component ratings ranged between .3 and 1.1 points.

A Multivariate analysis of variance (MANOVA) indicated a significant main effect of assessee gender [$F(8,1037)=5.02; p<.001$)] which supported hypothesis 2a. Following the significant main effect of assessee gender, univariate analysis of variance was performed on each of the eight dependent variables (Table 17). Results indicated significant differences between male and female assessees in 6 of the 8 components. As Table 17 indicates, significant differences were found in components entitled Teacher Plans

Effectively for Instruction, Maintains Environment, Manages Learner Behavior, Delivers Instruction, Provides for Student Involvement, and Assesses Student Progress.

Research Question 3: Does the gender of the assessor and the assessee interact to influence the performance ratings of teaching effectiveness? (That is, do same-sex evaluation ratings differ from opposite-sex evaluation ratings?)

Hypothesis 3a: There is a statistically significant assessor/assessee gender interaction effect.

A Multivariate analysis of variance (MANOVA) did not reveal a significant interaction effect of assessor gender by assessee gender [$F(8,1037)=1.07;p<0.380$)] (Table 17). However, the pattern of mean component ratings illustrated in Table 16 indicated that same-sex evaluation ratings differed slightly from opposite-sex evaluation ratings.Under female assessor conditions, same-sex evaluation ratings (female assessors with female assessees) were <u>higher</u> than opposite-sex evaluation ratings (female assessors with male assessees). Yet, under male assessor conditions, same-sex evaluation ratings (male assessors with male assessees) were <u>lower</u> than opposite-sex evaluation ratings (male assessors with female assessees). However, across both assessor gender conditions (male and female) and in general, same-sex evaluations had a lower overall scale mean rating of 8.43, whereas opposite-sex evaluations had a slightly higher overall scale mean rating of 8.52.

Table 17

Analysis of Variance for Assessor/Assessee Gender

	Assessee Gender	Assessor Gender	Interaction
Multivariate Analysis	**5.02***	.646	1.07
Univariate Analysis			
1.Plans Effectively	**16.66 ***	1.83	.126
2.Maintains Environment	**15.80***	.474	1.05
3.Maximizes Time	2.04	.285	2.02
4.Manages Behavior	**9.74***	.176	.831
5.Delivers Instruction	**15.44***	.805	.647
6.Presents Content	1.88	.016	1.12
7.Student Involvement	**20.21***	.527	1.36
8.Student Progress	**17.70***	.044	.266

***p<.001**

These slight differences in means, however, were not statistically significant.

Issue II: Ethnicity Bias

Research Question 4: Do African American and Caucasian assessors differ in their average ratings of assessees?

Hypothesis 4a: There is a statistically significant difference between the mean ratings given by African American and Caucasian assessors.

As Table 18 indicates, African American and Caucasian assessors differed in their overall mean ratings of assessees. African American assessors gave an overall scale mean rating of 8.63, and Caucasian assessors gave an overall mean component rating of 8.46. African American assessors gave the highest ratings on 7 of the 8 mean component ratings and Caucasian assessors gave the highest ratings on 1 of the 8 component ratings. A multivariate analysis of variance (MANOVA) revealed a significant main effect of assessor ethnicity [$F(8,1037)=1.94;p<0.051$)] (Table 19). This statistically significant difference in the ratings given by African American and Caucasian assessors supported hypothesis 4a.

Following the significant multivariate test of assessor ethnicity, a univariate analysis of variance was performed on each of the eight dependent variables (Table 18). Results indicated significant differences between African American

and Caucasian assessors in only 1 of the 8 components (Teacher Maintains Environment).

Research Question 5: Is there a difference in the average ratings of African American and Caucasian assessees?

Hypothesis 5a: There is a statistically significant difference in the mean ratings received by African American and Caucasian assessees.

As Table 18 indicates, there was a difference in the average ratings of African American and Caucasian assessees although the difference was small. On the average, the component ratings were consistently higher for Caucasian assessees, as compared to African American assessees.

Caucasian assessees had an overall scale mean rating of 8.72, whereas African American assessees had an overall scale mean rating of 8.37.

When considering the eight mean component ratings individually, Caucasian assessees scored higher than African American assessees in 7 of the 8 components. This trend was present regardless of assessor ethnicity. The difference between the average component ratings of Caucasian assessees and African American assessees ranged between 0.2 and 0.9 points.

A multivariate analysis of variance (MANOVA) indicated a marginally significant main effect of assessee ethnicity [$F(8,1037)=1.90; p<0.055$)] which tentatively supported hypothesis 5a. Following this multivariate test of main

Table 18

Mean Component Ratings (Assessor/Assessee Ethnicity)

		Assessor Ethnicity					
		White			Black		
Assessee Ethnicity		**White**	**Black**	**All**	**White**	**Black**	**All**
Plans Effectively	**X**	**12.30**	**11.65**	**12.3**	**12.15**	**11.82**	**12.06**
	S	2.18	2.29	2.19	2.10	1.94	2.07
Maintains Environment	**X**	**5.26**	**4.98**	**5.24**	**5.35**	**5.37**	**5.36**
	S	.906	1.05	.990	.819	.867	.830
Maximizes Time	**X**	**4.99**	**4.68**	**4.97**	**5.00**	**4.91**	**4.98**
	S	1.06	1.20	1.07	1.06	.995	1.04
Manages Behavior	**X**	**4.97**	**5.00**	**4.97**	**5.04**	**4.87**	**5.00**
	S	1.09	1.03	1.09	.994	1.04	1.00
Delivers Instruction	**X**	**9.89**	**9.22**	**9.85**	**9.96**	**9.62**	**9.87**
	S	1.90	1.92	1.91	1.70	1.84	1.77
Presents Content	**X**	**10.10**	**9.27**	**10.0**	**10.18**	**9.84**	**10.09**
	S	1.78	2.03	1.81	1.68	1.76	1.71
Student Involvement	**X**	**9.64**	**9.17**	**9.60**	**9.72**	**9.50**	**9.67**
	S	1.85	1.87	1.89	1.69	1.80	1.72
Student Progress	**X**	**12.38**	**11.89**	**12.3**	**12.56**	**12.00**	**12.42**
	S	2.20	2.13	2.20	1.98	2.02	1.99
Overall Scale Mean	**X**	**8.69**	**8.23**	**8.46**	**8.75**	**8.50**	**8.63**
	S	**1.33**	**1.40**	**1.27**	**1.21**	**1.22**	**1.15**

effect, a univariate analysis of variance was performed on each of the eight dependent variables (Table 19).

Results indicated significant differences between Caucasian and African American assessees in 4 of the 8 components, and marginally significant differences in another 2 components. As Table 19 indicated, significant differences were found in components entitled", Teacher Plans Effectively for Instruction, Maximizes Time, Delivers Instruction, Presents Content, Provides for Student Involvement, and Assesses Student Progress."

Research Question 6: Does the ethnicity of both the assessor and assessee interact to influence the performance ratings of teacher effectiveness? (That is, do same-ethnicity evaluation ratings differ from opposite-ethnicity evaluation ratings?)

Hypothesis 6a: There is a statistically significant assessor/assessee ethnicity interaction effect.

The pattern of mean component ratings illustrated in Table 18 indicated that same-ethnicity evaluation ratings differed slightly from opposite-ethnicity evaluation ratings. Under Caucasian assessor conditions, same-ethnicity evaluation ratings (Caucasian assessors with Caucasian assessees) were <u>higher</u> than opposite-ethnicity evaluation ratings (Caucasian assessors with African American assessees). Yet, under African American assessor conditions, same-ethnicity evaluation ratings (African American assessors with African American assessees) were <u>lower</u> than

Table 19

Analysis of Variance for Assessor/Assessee Ethnicity

	Assessee Ethnicity	Assessor Ethnicity	Interaction
Multivariate Analysis			
Univariate Analysis	**1.91***	**1.94***	1.40
1.Plans Effectively	**5.00***	.057	.787
2.Maintains Environment	2.24	**7.08***	2.67
3.Maximizes Time	**3.76***	1.13	.054
4.Manages Behavior	.401	.054	.959
5.Delivers Instruction	**7.40***	1.69	.753
6.Presents Content	**11.00***	3.38	1.88
7.Student Involvement	**3.63***	1.31	.475
8.Student Progress	**6.19***	.492	.023

***p<.001**

opposite-ethnicity evaluation ratings (African American assessors with Caucasian assessees). Overall scale mean component ratings indicated that African American assessors evaluating Caucasian assessees gave the highest scores (8.75), followed by Caucasian assessors evaluating Caucasian assessees (8.69). Lower scores were given when African American assessors evaluated African American assessees (8.50) and Caucasian assessors evaluated African American assessees (8.23).

Across both assessors ethnicity conditions (Caucasian and African American) and in general, same-ethnicity evaluations had a higher overall scale mean rating of 8.60, whereas opposite-ethnicity evaluations had a lower overall scale mean rating of 8.50. These slight differences in means, however, were not statistically significant. A multivariate analysis of variance (MANOVA) did not reveal a significant interaction effect of assessor ethnicity by assessee ethnicity [$F(8,1037)=1.39;p<0.193$)] (Table 19).

Issue III: Role Perception Bias

Research Question 7: Is there a difference in the average ratings given by the three types of assessors (principal, master teacher, and external assessors)?

Hypothesis 7a: There is a statistically significant difference between the mean ratings given by the three types of assessors (principal, master teacher, and external assessor).

Table 20

Mean Component Ratings Across Assessor Role

Assessor Role	Principal		Experienced Teacher		External Assessor	
	X	SD	X	SD	X	SD
Plans Effectively	12.10	2.12	12.29	2.17	12.15	2.24
Maintains Environment	5.32	.840	5.28	.891	5.12	1.03
Maximizes Time	5.01	.990	5.02	1.07	4.83	1.18
Manages Behavior	5.03	1.03	4.95	1.06	4.92	1.13
Delivers Instruction	9.91	1.85	9.89	1.86	9.58	1.98
Presents Content	10.02	1.77	10.16	1.77	9.84	1.85
Student Involvement	9.68	1.76	9.60	1.82	9.53	1.95
Student Progress	12.43	2.07	12.36	2.18	12.20	2.25
Overall Scale Mean	8.68	1.25	8.69	1.30	8.52	1.43

As Table 20 presents, the three types of assessors differed in their overall scale mean ratings of assessees, although the differences were small. Master teachers gave an overall scale mean rating of 8.69, followed by principals with a rating of 8.68 and external assessors with a rating of 8.52. Principals gave the highest ratings on 5 of the 8 mean component ratings (Teacher Delivers Instruction, Manages Learner Behavior, Delivers Instruction, Provides for Student Involvement, and Assesses Student Progress) and Master Teachers gave the highest ratings on 3 of the 8 component ratings (Teacher Plans Effectively for Instruction, Maximizes Time, and Presents Content). External assessors consistently gave the lowest ratings across all of the eight components.

A multivariate analysis of variance (MANOVA) revealed a significant main effect of assessor type [$F(8,1037)=1.83;p<0.023$] (Table 21), supporting hypothesis 7a.

Following the significant multivariate main effect of assessor type, a univariate analysis of variance was performed on each of the eight dependent variables (Table 21). Results indicated significant differences among principals, master teachers, and external assessors in only 1 of the 8 components. As Table 21 indicates, a significant difference was found in the component entitled "Teacher Maintains Environment."

Table 21

Analysis of Variance for Assessor Role

	Assessor Role
Univariate Analysis	1.83*
1.Plans Effectively	.847
2.Maintains Environment	3.36*
3.Maximizes Time	2.33
4.Manages Behavior	.947
5.Delivers Instruction	2.24
6.Presents Content	2.27
7.Student Involvement	.457
8.Student Progress	.748

*p.< .05

Research Question 8: If there is a difference in average ratings given by the three types of assessors, do principals give higher average ratings than master teachers and external assessors?

Hypothesis 8a: Principals are more lenient in rating assessees than other raters', therefore, principals will have the highest mean ratings among the three types of assessors.

As Table 20 indicates, the three types of assessors appeared to differ in their overall scale mean ratings of assessees as well as in individual component ratings. However, the means indicated that principals did not give higher average ratings to assessees than master teachers. Master teachers gave slightly higher ratings than principals and had an overall scale mean rating of 8.69. Principals followed with a rating of 8.68 and external assessors with a rating of 8.52.

Discussion for Study I

As the reliability estimates of internal consistency (Cronbach's alpha) show, the assessment instrument was moderately to highly reliable.

The results from data analyses addressing Issue I indicated that the differences in assessment ratings were attributable to the assessee gender, and not to the gender of the assessor or interaction of the assessee's and the assessor's gender. Female assessees had consistently higher

ratings than male assessees across every component regardless of the assessor gender. We now know that female assessees are rated higher, but why? Are female assessees actually better teachers than male assessees? Or could it be that females, rather than males, are more traditionally viewed as classroom teachers?

The results from data analyses addressing Issue II indicated that the differences in assessment ratings are attributable to assessee ethnicity and to assessor ethnicity. Caucasian assessees had consistently higher ratings than African American assessees across every component regardless of the assessor ethnicity. African American assessors consistently gave higher assessment ratings regardless of the assessee's ethnicity.

The results from data analyses addressing Issue III indicated that the differences in assessment ratings are attributable to the role assumed by the assessor. However, role differences were not found across every dependent variable. This makes interpreting role differences rather difficult. While there was a statistically significant difference in assessment ratings given by the three types of assessors, it was not clear with which role the difference exists. Individual component ratings appeared to indicate that those assessors assuming the role of principal gave higher ratings across 5 of the 8 components (63%), yet master

teachers had a slightly higher overall scale mean component rating.

One of the problems associated with Study I was that the independent variables (e.g. assessee gender) could not be actively manipulated. As such, there could be no case-effect link established but only associations between performance ratings and certain independent variables. Because of the conditions in Study I, it was very difficult to determine if actual differences in assessment ratings were attributable to actual differences in ability or bias. Study II provided the additional control needed, as it allowed certain assessment situations to be manipulated as Study I could not.

Results for Study II

Issue I: Gender Bias

Research Question 1. Do male and female assessors differ in their overall ratings of assessees?

Hypotheses 1a. Female assessors will rate both male and female assessees higher than male assessors.

Table 22 presents the mean rating of the four groups. As the table indicates, female assessors gave a mean effectiveness rating of 2.04 and male assessors gave a mean effectiveness rating of 1.98. Male and female assessors differed in their overall mean ratings of assessees, although the difference was small.

The mean effectiveness ratings presented in Table 22 supported hypothesis 1a, demonstrating that female assessors

rated both male and female assessors higher than male assessors. The mean effectiveness ratings given by female assessors to male assessees was 1.97 and to female assessees 2.13. The mean effectiveness ratings given by male assessors to male assessees was 1.96 and to female assessees 2.00.

It was apparent that female assessees were rated highest when assessed by male assessors. However, among male assessees, those who were rated by female assessors scored slightly higher than those who were rated by male assessors.

An analysis of variance (ANOVA) (2X2X2X2X3) did not reveal a significant main effect of assessor gender [$F(1,198)=0.98; p<.323$)] (Table 23); therefore, Hypothesis 1a was not supported.

Research Question 2: Is there a difference in the average ratings of male and female assessees?

Hypothesis 2a: There is a statistically significant difference in the mean ratings received by male and female assessees.

As Table 22 indicates, there was a difference in the average ratings of male and female assessees, although the difference was small. On the average, the effectiveness ratings were consistently higher for female assessees, as compared to male assessees. Female assessees had a mean effectiveness rating of 2.07, whereas male assessees had a mean effectiveness rating of 1.96. Female assessees consistently scored higher than male assessees regardless of

Table 22

Effectiveness Ratings (Assessor/Assessee Gender)

	Assessor Gender					
	Female			Male		
Assessee Gender	Female	Male	All	Female	Male	All
Effectiveness X	**2.13**	**1.97**	**2.10**	**2.00**	**1.96**	**1.98**
S	**0.78**	**0.88**	**0.84**	**1.03**	**0.63**	**0.85**

Table 23

Multivariate Analysis of Variance

Source	F	Sig of F
Assessee Gender	0.80	.372
Assessor Gender	0.98	.323
Assessee Ethnicity	2.26	.135
Assessor Ethnicity	.00	.962
Assessee Gender X Assessor Gender	0.21	.647
Assessee Ethnicity X Assessor Ethnicity	2.57	.111
4-Way Interaction	8.73	**.01***

***p<.05**

the assessor's gender. However, the difference between male assessee and female assessee mean effectiveness ratings was only slight, being approximately 1.1 points.

An analysis of variance (ANOVA) did not indicate a significant main effect of assessee gender [F(1,198)=0.80;p<0.372)] and therefore did not support Hypothesis 2a. Research Question 3: Do the genders of the assessor and the assessee interact to influence the performance ratings of teaching effectiveness? (That is, do same-sex evaluation ratings differ from opposite-sex evaluation ratings?)

> Hypothesis 3a: There is a statistically significant assessor/assessee gender interaction effect.

The pattern of mean component ratings illustrated in Table 22 indicated that same-sex evaluation ratings differed slightly from opposite-sex evaluation ratings. Under female assessor conditions, same-sex evaluation ratings (female assessors with female assessees) were <u>higher</u> than opposite-sex evaluation ratings (female assessors with male assessees). Yet, under male assessor conditions, same-sex evaluation ratings (male assessors with male assessees) were <u>lower</u> than opposite-sex evaluation ratings (male assessors with female assessees). However, across both assessor gender conditions (male and female) and, in general, same-sex evaluations had a higher mean effectiveness rating of 2.05 whereas opposite-sex evaluations had a lower mean

effectiveness rating of 1.99. These slight differences in means, however, were not statistically significant.

Mean effectiveness ratings indicated that female assessors evaluating female assessees gave the highest scores (2.13), followed by male assessors evaluating female assessees (2.00). Lower scores were given when female assessors evaluate male assessees (1.97) and when male assessors evaluated male assessees (1.96).

An analysis of variance (ANOVA) did not reveal a significant double interaction effect of assessor gender by assessee gender [$F(1,198)=0.241;p<0.647$] (Table 23). However, since the four-way interaction was significant, these results should be interpreted cautiously. Testing the gender interaction effect within assessee ethnicity found no effect for Caucasian assessees [$F(1,198)=0.05;p<0.8180$] or for African American assessees [$F(1,198)=1.21;p<0.2720$]. A simple 2X2 ANOVA, similar to that which was used in Study I did not result in a significant gender interaction effect [$F(1,213)=0.236;p<0.6270$]

Issue II: Ethnicity Bias

Research Question 4: Do African American and Caucasian assessors differ in their average ratings of assessees?

Hypothesis 4a: There is a statistically significant difference between the mean ratings given by African American and Caucasian assessors.

As Table 24 indicates, African American and Caucasian assessors differed only slightly in their mean effectiveness ratings of assessees. African American assessors gave a mean effectiveness rating of 2.03, and Caucasian assessors gave a mean effectiveness rating of 2.00. However, an analysis of variance (ANOVA) did not reveal a significant main effect of assessor ethnicity [F(1,198)=.000] (Table 23).

Research Question 5: Is there a difference in the average ratings of African American and Caucasian assessees?

Hypothesis 5a: There is a statistically significant difference in the mean ratings received by African American and Caucasian assessees.

As Table 24 indicates, there was a difference in the average ratings of African American and Caucasian assessees although the difference was small. On the average, effectiveness ratings were consistently higher for Caucasian assessees, as compared to African American assessees. Caucasian assessees had a mean effectiveness rating of 2.11, whereas African American assessees had a mean effectiveness rating of 1.91. This trend was present regardless of assessor ethnicity. The difference between the effectiveness ratings of Caucasian assessees and African American assessees was only 0.2 of a point.

An analysis of variance (ANOVA) did not reveal a significant main effect of assessee ethnicity [(F (1,198)=2.26;p<.135)] which did not support hypothesis 5a.

Table 24

Effectiveness Ratings(Assessor/Assessee Ethnicity)

	Assessor Ethnicity					
	White			Black		
Assessee Ethnicity	White	Black	All	White	Black	All
Effectiveness X	**2.18**	**1.80**	**1.99**	**2.00**	**2.07**	**2.04**
S	**0.88**	**0.81**	**0.87**	**0.82**	**0.82**	**0.82**

Research Question 6: Does the ethnicity of both the assessor and assessee interact to influence the performance ratings of teacher effectiveness? (That is, do same-ethnicity evaluation ratings differ from opposite-ethnicity evaluation ratings?)

Hypothesis 6a: There is a statistically significant assessor/assessee ethnicity interaction effect.

The pattern of mean component ratings illustrated in Table 24 indicated that same-ethnicity evaluation ratings differed slightly from opposite-ethnicity evaluation ratings. Under Caucasian assessor conditions, same-ethnicity evaluation ratings (Caucasian assessors with Caucasian assessees) were higher than opposite-ethnicity evaluation ratings (Caucasian assessors with African American assessees). Under African American assessor conditions, same-ethnicity evaluation ratings (African American assessors with African American assessees) were also higher than opposite-ethnicity evaluation ratings (African American assessors with Caucasian assessees). Overall mean effectiveness ratings indicated that Caucasian assessors evaluating Caucasian assessees gave the highest scores (2.18), followed by African American assessors evaluating African American assessees (2.07). Lower scores were given when African American assessors evaluated Caucasian assessees (2.00) and Caucasian assessors evaluated African American assessees (1.80).

Across both assessors' ethnicity conditions (Caucasian and African American) and same-ethnicity evaluations had a higher mean effectiveness rating of 2.13 whereas opposite-ethnicity evaluations had a lower mean effectiveness rating of 1.90. While there were apparent differences in means, they were statistically not significant. Results of an analysis of variance (ANOVA) did not support a significant double interaction effect of assessor ethnicity by assessee ethnicity [$F(1,198)=2.57;p<0.11$)].

In addition to the two-way interaction effects tested within the two-factor ANOVA Design, A four-way interaction effect also was tested in a full-factor ANOVA design. The highest order interaction effect (Assessee Gender by Assessor Gender by Assessee Ethnicity by Assessor Ethnicity) was significant $F(1,198)=8.73;p<0.01)$. When a significant four-way interaction effect in the full-factor ANOVA design was found, tests of simple effects were conducted. Testing the ethnicity interaction effect within assessee and assessor gender found significant effect for female assessors [$F(1,198)=6.73;p<0.01)$] and also for male assessees [$F(1,198)=9.39;p<0.01)$] (Table 23).

Supplemental data analyses using a traditional two factor ANOVA similar to the one used in Study I was conducted. The traditional two factor ANOVA revealed a significant (Assessee Ethnicity by Assessor Ethnicity) interaction effect [$F(1,3)=3.79;p<0.053$]. In the full four

factor ANOVA, this interaction effect was found in male assessees and female assessors separately.

Discussion for Study II

Utilizing a strictly experimental design in Study II, the effect of gender and ethnicity on the assessment process was clearly determined. While the gender, ethnicity, and ability level of the teacher were held constant, assessors still differed in their ratings of the teacher. While the differences in the assessment ratings were not statistically significant they were of practical relevance. Having removed all other confounding elements, differences in assessment ratings were the result of systematic errors in measurement and observation is "bias". In Chapter 1 of this paper, bias was described as an inclination or preference that interferes with impartial judgement. This preference was noticeable when the results of Study II were considered, but was not so easily recognizable within Study I.

The results of this experimental study were quite interesting. While the findings from data analyses addressing Issue I did not show that the differences in effectiveness were attributable to the gender of the assessor or assessee nor to the interaction of both the assessee's and the assessor's, gender, there were some notable observations.

While there were no significant effects related to gender, the pattern of effectiveness ratings was confirmatory. Consistent with Study I, female assessors gave

higher ratings than male assessors, and female assessees received higher ratings than male assessees.

The hierarchy of mean ratings among male and female assessors and assessees in Study II was an exact replica of those in Study I. As in Study I, mean effectiveness ratings indicated that female assessors evaluating female assessees gave the highest scores, followed by male assessors evaluating female assessees. Lower ratings followed with female assessors evaluating male assessees and male assessors evaluating male assessees.

Another gender-specific pattern that is worthy of note was that of same-sex and opposite-sex evaluations. Confirming what was revealed in Study I, Study II also produced higher same-sex evaluations and lower opposite-sex evaluations. In fact, in Study I, under female-assessor conditions, same-sex evaluation ratings were <u>higher</u> than opposite-sex evaluation ratings and under male assessor conditions, same-sex evaluation ratings were <u>lower</u> than opposite-sex evaluation ratings. Study I findings were confirmed by Study II.

The findings from data analyses addressing Issue II did not conclude that the differences in effectiveness were attributable to the ethnicity of the assessor or assessee but rather to the interaction of both the assessee's and the assessor's ethnicity.

As was found in Study I, Study II revealed that Caucasian assessees receive higher ratings than African

American assessees and African American assessors gave higher ratings than Caucasian assessors. Again, it is satisfactory to find consistent confirmatory results. While the ethnicity-specific patterns such as same-ethnicity and opposite-ethnicity evaluation patterns were slightly different than those found in Study I, they were still important. Study II indicated that Caucasian assessors evaluating Caucasian assessees gave the highest ratings, followed by African American assessors evaluating African American assesses. This pattern of results clearly showed that higher effectiveness ratings were being given to assessees who were similar in ethnicity to the assessor. This finding was indicative of "intergroup theory" or "positivity bias theory," which was discussed in Chapter 2 of this paper. These theories provided the rationale that higher ratings were given for same group ratees.

In conclusion, the pattern of results found in Study II extend and confirm those found in Study I. Assessors and assessees of differing gender and ethnicity were consistently behaving in the same manner. However, due to the smaller sample size, cell size, and limited number of dependent variables in Study II, meaningful, statistically significant differences were not revealed as they were in Study I.

In summary, it is assessee gender, assessee ethnicity, assessor ethnicity, and assessor role which contributed to differences in assessment ratings. While the mean differences

appeared to be only slight, they were large enough to produce statistically significant differences. These findings were important in that they supported the idea of bias within assessment systems. Differences in assessment ratings resulting from one's gender, ethnicity, and role may be the direct result of biases. It is clear that there was strong support for ethnic bias, gender bias, and role bias within the Louisiana Teacher Assessment Program.

CHAPTER 5: CONCLUSIONS

The reform movement in teacher education continues to focus upon stringent performance evaluation and appraisal systems. Demand for teacher accountability, coupled with powerful political initiatives, has led to the development and legislatively enacted teacher-assessment systems. It is quite apparent that states and local districts are diligently working to devise and implement methods and programs of teacher evaluation. However, it is questionable whether or not these teacher evaluation systems are providing unbiased information.

As has been mentioned in previous chapters, any system which relies on observation as a primary mechanism for rating performance may be affected by the limits inherent in observational methods. Bias is one of the most obvious and limiting phenomena associated with observational techniques and is quite difficult to control. The concept of bias, particularly as a possible contaminator of assessment ratings utilized in performance evaluation, was the focus of this research as well.

This study addressed the notion of performance evaluation as a "type of social perception";therefore it inevitably entails "forming beliefs about the quality of a person's task performance based upon perceptions of the person's activities (Foschi & Lawler, 1994)."

Unfortunately, our reliance on human instruments and their perceptions of performance may lead to biases. These perceptual biases, in turn, may affect performance evaluations. Perceptual biases involving social characteristics such as gender, ethnicity, and role are among these.

Two studies were done to explore the possibility of such bias. One was an ex-post-facto study of teacher-evaluation results in Louisiana. The other was an experimental study with the same group of evaluators as in the first study, but with an experimentally manipulated evaluation situation.

Study II was important for many reasons. First, the experimental design provided the additional control needed to actively manipulate independent variables to strengthen the cause-and-effect relationship. Under the conditions of Study I, it was difficult to determine if differences in assessment ratings were attributable to actual differences in ability and performance or to bias. Study II utilized experimental teaching situations, which did not occur normally, in the pilot; therefore, it provided a sound alternative to analyzing assessment ratings for subgroups of the population which was not available in Study I (i.e., black males). Secondly, it enabled effectiveness ratings to be analyzed for specific combinations of assessors and assesses which were not so easy to determine in Study I.

The results of both studies confirmed the existence of some bias in the assessment process. Statistically significant differences were noted in performance evaluation attributable to assessee gender, assessee ethnicity, assessor ethnicity, and assessee role. Assessment data revealed that female assessees were consistently rated higher than male assessees and that female assessors consistently rated assessees higher than male assessors. However, male assessors were found to give higher ratings to female assessees than female assessors. The presence of gender bias is not particularly surprising, as an extensive review of the literature reveals that it occurs within many disciplines and environments. However, there has been very little empirical research documenting gender bias within classroom contexts, which adds to the value of these findings. In accordance with the findings of Feldman (1983); Basow and Distenfield (1985); Basow and Howe (1987) females were consistently rated higher than males. As in many other studies, the findings from Study I support the unfortunate reality of gender bias.

Assessment data also revealed that Caucasian assessees were consistently rated higher than African American assessees and that African American assessors consistently rated assessees higher than Caucasian assessors. Again, when considering the ethnicity of the assessor in conjunction with the ethnicity of the assessee, African American assessors

were found to give higher ratings to Caucasian assessees than Caucasian assessors did. Finding this second form of bias supports previous research.

While bias in any form is bad, ethnicity bias is one of the more sensitive and highly publicized issues affecting employment opportunities. Unlike gender bias, ethnicity bias and its relationship to job performance evaluation has been thoroughly documented (Greenhaus and Parasuraman, 1993; Martocchio and Whitener, 1992; Kraiger and Ford, 1990).Research addressing issues of ethnicity and bias in teacher assessment and evaluation is limited, which, again, adds to the value of these findings.

As has been mentioned, the effect of assessor's role has been revealed. Although the differences were small, results indicated that the three types of assessors differed in their performance ratings of assessees. Supporting previous research (Cronin and Capie, 1986; Ellett and Capie, 1985; Ellett, Teddlie and Niak, 1991; Kelly, 1985), principals and other on-campus evaluators (master teachers) gave higher performance ratings than off-campus (external assessors) evaluators. This is a very important finding because it clearly supports the concerns of fellow researchers that closer examination of the role a person serving as a performance assessor is needed (Ellett & Capie, 1985; Acheson, Smith & Stuart, 1986; Garland, 1989).

Statistical Significance and Practical Significance

Although statistically significant differences in assessment ratings were found in Study I, the magnitude of effects was relatively small. Small differences in assessment means may have little or no practical significance even though they may be significant in the statistical sense. What is of importance is the pattern of findings from both studies. In both studies the pattern of assessment ratings for all three issues was consistent. While Study IT failed to reveal any statistically significant effects, the consistency of the findings was the connecting link.

When considering both studies, it is possible that intensive training of observers reduced the bias due to gender and/or ethnicity. However, the fact that some residual degree of bias was <u>still</u> present points to the methodological and practical need for caution when using observation. It is methodologically and practically important that an examination of the possibility of gender, ethnicity, or role biases be conducted because they may distort assessment results.

If large-scale state-wide appraisal systems are implementing costly teacher-assessment programs, consideration must be given to the most cost-effective approach. If the intent is to establish and maintain a valid teacher assessment program, then issues of bias must be addressed prior to implementation. If a seven-day rigorous

assessor training program is devised for purposes of attaining assessor reliability, than what does that say about the training components when bias in assessment ratings still remains?

It is quite evident that even well-trained assessors had slight biases while conducting performance assessment. Given the intent of the assessor training, bias is still active in the assessment process. It seems no amount of "bias awareness" training can fully remove inherent biases that individuals bring to performance assessment situations.

The findings and significance of both studies are of importance when considered separately as well as jointly. When considered separately, they are methodologically sound and comprehensive works which strengthen and add value to the limited empirical research which has been conducted in classroom contexts. When considered as a joint effort, these studies offer a solid new approach to addressing issues of gender, ethnicity, and role biases within educational environments. While the assessment literature reveals that there have been a number of studies conducted over the past twenty years, these studies provide only small pieces of a much larger puzzle. This research effort systematically incorporates gender, ethnicity, and role effects, using more that one methodological approach (e.g., causal comparative and experimental) and more than one mode of inquiry (e.g.,

observational and survey). These studies have also used more than one mode of instrumentation, analysis, and sample size.

Implications

While the magnitude of effects were rather small, these results still point to potential bias within the assessment process. Given the confines of the studies (i.e., limited sample and limited observations), what do these findings mean in a larger context? Are we to assume that all teacher-assessment systems are biased or that all assessors are biased? Are we to accept these biases, tolerating their existence as an unavoidable circumstance?

Given the many differing teacher assessment systems across the country, it would be rather difficult to make any generalizations based on these studies. Not only will the conditions and limitations of each study vary, so will the true teaching ability of the assessee.

To complicate matters, assessment is only a "perception" or "observation" of a performance or ability. Based on the Classical Theory of Reliability (Crocker & Algina, 1990) the "observed" or "perceived" score is made up of the applicant's true score plus any error. We will never know an assessee's "true" score, only the "observed" score. When we add error, such as measurement error, to the true score then add known biases into the error component, the observed score becomes even more distorted.

When bias exists and true measures of performance do not, the likelihood that accurate inferences will be made decreases. For instance, females assessees received higher ratings possibly because they were just better teachers than males; therefore, the issue of bias in assessment ratings become even more difficult to assess.

One of the more important aspects is that differences were small and, therefore, the resulting biases may be minimal and tolerable. A possible implication may be that given more observations, these particular differences may level out, thereby minimizing, if not eliminating, any biases.

Another implication resulting from this study may be that perhaps bias may not have been found in Study I, if the assessor assignment had been random. Unfortunately, the establishment of the assessment team varies, as the procedure is sometimes based on immediate needs, availability, and local politics. When assessors were randomly assigned in Study II, there were only slight biases in assessment ratings.

Another implication may be the number of assessors assigned to any one assessment team. It may be that an increase in the number of assessors, and therefore observations, may affect the assessment ratings by minimizing differences. The number of assessors assigned to an assessment team was a politically motivated decision, as were

many other decisions, rather than a psychometrically motivated decision.

One of the possible ways to reduce the biases may be not to rely strictly on observational measures. Perhaps more accountable measures need to be explored. It may be that more objective, less intrusive, methods such as video cameras and observational coding schedules or self-report measures need to be explored. Perhaps class observations can be used in conjunction with other methods. A multi-method approach would create a more realistic idea of the teacher's ability. As the criticism has been made, six one-hour, pre-announced classroom visits may not be adequate. Essentially what the assessor is getting is a controlled "snap-shot" of the classroom teacher's ability. In addition, with scheduled preannounced visits, the teacher may have rehearsed the lesson enough to give a polished performance or a good show, rather than a "typical" teaching demonstration.

Both studies clearly illustrated the need to more closely examine our current teacher-assessment systems. All local and state teacher assessment systems should be extensively reviewed for bias periodically, but checking for bias is not enough. How bias, if any, affects outcome measures and the implications that follow will need to be fully explored. In addition, alternative measures of decreasing or resolving bias in teacher-assessment systems will need to be developed and implemented.

As the nation moves toward teacher-assessment systems that rely on observational rating performances, one must be prepared to extrapolate true assessment ratings from those that are confounded by biases. Differences in assessment ratings are tolerable, but not if they are the result of gender, ethnicity, or role biases rather than true differences in assessees performance.

General Limitations of the Studies

Study I

The major limitation of Study I is within the assessment data. The LTAI used a three-point rating scale. As such, the ratings had a small range and a small variance. The nature of the assessment system is to decrease any naturally occurring variance. Variance is decreased by limiting the range of the rating scale, the number of assessors, the number of observations, and by encouraging the assessors to utilize a uniform rationalized rating scale.

Another limitation that is related to assessment ratings is the manner in which assessment teams were composed. Although there was a uniform nomination process by which assessors were selected, how they were assigned to the assessee's team varies. It is possible that the manner or process by which teams are comprised may be biased in itself, leading to bias within assessment ratings.

Another limitation of Study I was that only one year (two semesters) of assessment data was available to analyze.

Despite the fact that these were the only data that actually impacted certification decisions, a longer period of observation would yield more reliable data. By analyzing only one semester of data, the number of available observations and assessment ratings was limited and the ability to generalize the results was also limited. Having only one semester of data to analyze, it was difficult to determine if any bias in assessment that may have occurred was an isolated event or if it was a part of a pattern within the assessment process. The patterns that have been established in the studies are important to future analysis.

Another limitation which was related to the pilot sample was the willingness of the LEAS and its personnel participating in the pilot. While a sub-sample of 20 LEAS agreed to participate in the pilot, it might be that these LEAs greatly differed from the other LEAs who did not agree to participate. Since participating in the pilot was voluntary, perhaps those LEAs and their personnel who agreed to participate were different in a way that would be essential to the findings of this study. It could be that those LEAs who volunteered to participate had excellent teachers who had no reluctance to being assessed, whereas the other LEAs had poor-performance teachers who were reluctant to being assessed.

Study II

While most of the participants from the pilot (1993-1994) continued assessing during implementation (1994-1995), those who continued to assess were required to attend a three day update training session. The three-day update training session re-oriented the assessors to the assessment process and the instruments, and, more importantly, extensively reviewed the assessment criteria, LCET. No data is available for assessors for Study II who received the additional assessor training.

Another possible limitation may be the generalizability of the findings from Study II. Study I used actual assessment results from authentic classroom observations. Study II used assessment results in response to a hypothetical scenario; the process of assessment might have been different across the two studies. In other words, there might be a discrepancy in what ratings of effectiveness assessors actually gave and what they "think" they gave.

A final limitation could be the sample size. Given that Study II had only 214 observations covering 16 experimental stimuli conditions, the actual cell sizes were moderately small. A larger sample might have produced different findings. It may be that, given more African American assessors, the effects related to ethnicity would be different.

Given the above-mentioned limitations, the need for future modifications or improvement is evident. First, more repeated observations would greatly improve the credibility of the studies and the inferences made from the studies. While these studies focused on the pilot data, the next logical step would be to analyze the full-scale assessment data from the first year of implementation. The actual assessment program, rather than the pilot program, would provide the next researcher with a much larger and more representative sample with which to study bias.

Another area for improvement may lie in the actual methodology employed. Given more data to analyze, trend analysis may be quite feasible. How are the assessee's ratings over time and how does bias function over extended periods? It would be interesting to look at assessment differences over assessment semesters or even over years.

It may also be interesting to develop and validate a parallel form assessment instrument or perhaps identify variables that correlate or predict success on LTAI. Perhaps a teacher's college GPA or NTE score could be a valid predictor of performance on the LTAI. If so, researchers may want to use these measures in conjunction with assessment results in multivariate analysis to better determine the teacher's level of competence.

To replicate this study as it has been described in this paper would be beneficial but there is certainly room for

extension and expansion. It could be very important to analyze both local teacher evaluation data and state teacher evaluation data. Both assessment systems utilize the same LCET assessment criteria and the same rating scale but for different reasons. The local teacher-assessment program, on the other hand, is for continued employment, contract purposes, and teacher accountability purposes. The state teacher assessment program is purely for certification and licensing purposes. It would be interesting to see if state assessment results are consistent with local assessment results. Are teachers being consistently assessed, or does the reason behind the assessment affect the assessment ratings?

Another important issue to explore is teacher accountability. Since local teacher evaluation is for new as well as experienced teachers, how do their assessment ratings compare? Are experienced teachers' more competent than new teachers? Are the experienced teachers performances "up to par" given the length of time they have been out of curriculum and instruction courses when compared to the more recently educated new teachers? The issue of competence among teachers, regardless of age and experience, continues to be of interest.

Additional Questions

In designing these studies, many questions evolved. One of the most basic yet controversial questions is that of pre-

existing prejudice and/or bias. A baseline, general measure of prejudice and/or bias, prior to assessment would perhaps provide the researcher with a better framework to analyze the assessment results. If it is known how an assessor regards women or African Americans in general, prior to conducting teacher assessment, might we have a better understanding of what that assessors' observation ratings mean? While an instrument measuring prejudice and/or bias may seem rather obvious, intrusive, deceptive, or even inappropriate, it could very well shed some light on assessment ratings.

Another question relating to bias lies within the history of the teaching profession. Early education classes reveal that females have always been traditionally placed in teaching roles. Females were considered to be fit for teaching due to their care-giving abilities and maternal nurturing instincts. Males, however, were not. When considering the gender-related patterns from both studies, perhaps this historic concept of gender and teaching plays an important role. It is possible that female assessees received higher assessment ratings from both male and female assessors because the assessors viewed female teachers as being naturally better teachers.

Another question directly related to assessment ratings is the order of assessment. Could the order in which an assessor conducted an assessment affect ratings? If it follows that principals and master teachers are more lenient

evaluators and they rate last, are their ratings going to be higher than those who rate first? And if so, are these ratings a reflection of the assessees' natural progression towards improvement or the order of assessor? An order effect would be an interesting aspect to explore.

Another question of interest is that of objectivity with regard to bias. If objectively is regarded as a myth, (as Scriven (1988) does) then bias is to be expected. However, if we hold the idea of objectivity as being attainable, what can we do to ensure objectivity? Can we ensure objectivity through extensive awareness training? Is objectivity attained as a result of limited experiences, or is objectivity ensured through constant re-focus and effort? The possibility of maintaining objectivity when there is a reliance on people to perform as human instruments measuring performance remains questionable.

REFERENCES

Abramowitz, C.V., & Dokecki, P.R. (1977). The politics of clinical judgement: Early empirical returns. Psychological Bulletin, 84, 460-476.

Abramson, P.R., Goldberg, P.A., Greenberg, J.H., and Abramson L.M.(1977).The talking platypus phenomenon: Competency ratings as a function of sex and professional status. Psychology of Women Quarterly, 22, 124.

Acheson, K. & Smith, S. (1986). It is time for principals to share the responsibility for instructional leadership with others. OSCC-Bulletin, 29 (6).

AERA/APA/NCME.(1985).Standards for educational and psychological testing. Washington, DC: American Psychological Association.

Alderfer, C.P., Alderfer., Tucker, L., & Tucker, R. (1980). Diagnosing race relations in management. Journal of Applied Behavioral Science, 27, 135-166.

Al-Issa, J. (1980). The psychopathology of women. Englewood Cliffs, NJ: Prentice-Hall.

Arvey, R.D. (1979). Unfair discrimination in the employment interview: Legal and psychological aspects. Psychological Bulletin, 36, 736-765.

Arvey, R.D. (1986). Sex bias in job evaluation procedures. Personnel Psychology, 39, 315-335.

Association of Teacher Educators. (1988). Teacher Assessment. Reston, VA: Author.(ERIC Document Reproduction Service No.ED 289869)

Barnes, S. (1987). The development of the Texas Teacher Appraisal System. Paper presented at the annual meeting of the American Educational Research Association, Washington, D.C. (ERIC Document Reproduction Service No. ED 294323)

Basow, S.A., & Distenfield, S. (1985). Teacher expressiveness: More important for male teachers than female teachers. Journal of Educational Psychology, 77, 45-52.

Basow, S.A., & Howe, K.G. (1987). Evaluations of college professors: Effects of professors' sex-type and sex, and student's sex. Psychological Reports, 60, 671-678.

Bass, B.A., & Barrett, C.V. (1972). Man, work, and organization. Boston: Allyn and Bacon.

Beauvais, C. & Spence, J.T. (1987). Gender, prejudice, and categorization. Sex Roles, 16, 89-100.

Bem, S. (1974). The measurement of psychological androgyny. Journal of Consulting and Clinical Psychology, 42, 155-162.

Bennett, S.K. (1982). Student perceptions and expectations for male and female instructors: Evidence relating to the question of gender bias in teaching evaluation. Journal of Educational Psychology, 74, 170-179.

Bigoness, W. (1976). Effect of applicant's sex, race, and performance on employer's performance ratings: Some additional findings. Journal of applied psychology, 61(1), 80-84.

Blackman, B.I. (1993). Qualpro user's manual: Version 4. Tallahassee, FL: Impulse Development Co.

Blumrosen, R.G. (1979). Wage discrimination, job segregation, and Title VII of the Civil Rights Act of 1964. University of Michigan Journal of Law Reform, 12, 397-502.

Borg, W., & Gall, M. (1989). Educational Research an introduction (3rd ed.) White Plains, NY: Longman.

Buczek, T.A. (1981). Sex biases in counseling. Counselor retention of the concerns of a female and a male client. Journal of Counseling Psychology, 28, 13-21.

Capie, W., Anderson, S.J., Johnson, C.E., & Ellett, C.D. (1980). Teacher Performance Assessment Instruments: A Handbook for Interpretation. Athens, GA: College of Education, University of Georgia.

Chauvin, S.W., & Ellett, C.D. (1991). Replacing lifetime certification with a renewable credential: A survey of Louisiana educators' perceptions of the Louisiana teaching internship and statewide teacher evaluation programs. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago. (ERIC Document Reproduction Service No. ED 335410).

Chauvin, S.W., Loup, K.S., & Ellett, C.D. (1991, April). Development and validation of a comprehensive assessment system for teaching and and learning. Paper presented at the Annual Meeting of the American Educational

Research Association, Chicago. (ERIC Document Reproduction Service No. ED 335410)

Chesler, P. (1972). Women and madness. Garden City: Doubleday.

Claudet, J. (1990). Review of pertinent literature and theoretical foundations of the Louisiana System for Teaching/Learning Assessment and Review. Technical report, Louisiana Teaching Internship and Statewide Teacher Evaluation Projects, College of Education, Louisiana State University, Baton Rouge, La.

Cobb, Terry (1984). The Illusion of Independence in Evaluation. North Carolina (ERIC Document Reproduction Service No. ED 292831)

Collier, H. (1982). Counseling women: A guide for therapists. New York: Macmillian

Collins, A. (1990). Novices, Experts, Veterans, and Masters: The Role of Content and Pedagogical Knowledge in Evaluating Teaching. Paper presented at the annual meeting of the American Educational Research Association, Boston. (ERIC Document Reproduction Service No. ED 319815)

Cronin, L.L. & Capie, W. (1986). The influence of daily variation in teacher performance on the reliability and validity of assessment data. Paper presented at the annual meeting of the American Educational Research Association, San Francisco. (ERIC Document Reproduction Service No. ED 274704)

Cooper, W.H. (1981). Ubiquitous halo. Psychological Bulletin, 90(2), 218-244.

Cooper, W.H. (1985). The effects of familiarity, gender, and institutional prestige on evaluative judgements of convention program proposals. Journal of Research and Development in Education. 18(3)

Darling-Hammond, L., Wise, A.E., & Pease, S.R. (1983). Teacher evaluation in the organizational context: A review of the literature. Review of educational research, 53(3), 285-327.

Defino, M., & Hoffman, J. (1986). A status report and content analysis of state mandated teacher induction programs. (Technical Report No.9057) Austin, TX: The University of Texas at Austin, Research and Development Center for Teacher Education.

Denzin, Norman K. (1978). The logic of naturalistic inquiry. In Sociological Methods: A Sourcebook. New York: McGraw-Hill.

Dukes, R., & Victoria, G. (1989). The effects of gender, status, and effective teaching on the evaluation of college instruction. Teaching Sociology, 17, 447-457.

Duckett, W. (1985). The competent evaluator of teaching: A CEDR monograph. (Report No. ISBN-0-87367-720-X). Bloomington, IN: Phi Delta Kappa, Center on Evaluation and Research. (ERIC Document Reproduction Service No. ED 266536)

Ellett, C.D. (1990). A new generation of classroom-based assessments of teaching and learning: Concepts, issues and controversies from pilots of the Louisiana STAR. Baton Rouge, LA: Teaching Internship and Statewide Teacher Evaluation Projects, College of Education, Louisiana State University.

Ellett, C.D. & Capie, W. (1985). Assessing meritorious teacher performance: A differential validity study. Paper presented at the annual meeting of the American Educational Research Association, Chicago. (ERIC Document Reproduction Service No. ED 264255)

Ellett, C.D., Garland, J. & Logan, C. (1987). Content classification, synthesis and verification of eight large-scale teacher performance assessment instruments. Research report, Teaching Internship Project, Baton Rouge, LA: College of Education, Louisiana State University, Baton Rouge, LA.

Ellett, C.D., Loup, K., & Chauvin, S. (1989). System for teaching and learning assessment and review (STAR). Statewide Teaching Internship and Teacher Evaluation Programs Form. College of Education, Louisiana State University, Baton Rouge, LA.

Ellett, C.D., Teddlie, C., & Niak, N. (1991). The effects of high stakes certification demands on the generalizability and dependability of a classroom-based teacher assessment system. Paper presented at the annual meeting of the American Educational Research Association, Chicago. (ERIC Document Reproduction Service No. ED 335408)

Etaugh, C., & Sanders, S. (1974). Evaluation of performance as a function of status and sex variables. Journal of Educational Psychology, 94, 237-241.

Feldman, K.A. (1983). Course characteristics and college teachers as related to evaluations they receive from students. Research in Higher Education, 18, 3-124.

Fernandez, J.P. (1981). Racism and sexism in corporate life: Changing values in American business. Lexington, MA: Lexington Books.

Fiske, D.W. (1978). Strategies for personality research. San Francisco: Jossey-Bass.

Foschi, M., & Lawler, E. (1994). Group Processes: Sociological Analyses. Chicago: Nelson-Hall Publishers.

Freedman, S.M., & Phillips, J.S. (1988). The changing nature of research on women at work. Journal of Management, 14, 231-251.

Garland, V. (1989). Cultivating excellence: A curriculum for excellence in school administration, III. Supervision of the administrative staff: The superintendent's role. New Hampshire, CT: University of New Hampshire, New Hampshire School Administrators Association.

Georgia Department of Education (1994). Athens, Georgia.

Goldberg, P.A. (1968). Are women prejudiced against women? Transaction, 5:28-30.

Gordon, R., & Owens, S. (1988). The effect of job level and amount of information on the evaluation of male and female job applicants. Journal of employment counseling. 25, 160-171.

Greenfield, William (1987). Instructional Leadership. Boston: Allyn and Bacon, Inc.

Greenhaus, J.H., & Parasuraman, S., (1993). Job performance attributions and career advancement prospects: An examination of gender and race effects. Organizational behavior and human decision processes, 55(2), 273-297.

Greenhaus, J.H., & Parasuraman, S., & Wormley, W.M. (1990). Effects of race on organizational experiences, job performance evaluations, and career outcomes. Academy of Management Journal, 33(1) 64-86.

Guion, R.M. (1965). Personnel Testing. New York: McGraw-Hill.

Hamner, W.C., Kim, J.S., Baird, L, & Bigoness, W.J. (1974). Race and sex as determinants of ratings by potential

employers in a simulated work-sampling task. Journal of Applied Psychology, 59(6), 705-711.

Harris, M.B. (1976). The effects of sex, sex-stereotyped descriptions and institution on evaluations of teachers. Sex Roles, 2,15-21.

Hastie, R. & Park, B. (1986). The relationship between memory and judgement depends on whether the judgement task is memory-based or on-line. Psychological Review, 93, 258-268.

Herman, J.L., Morris, L.L., & Fitz-Gibbon, C.T., (1987). Evaluator's Handbook. Newbury Park, CA: Sage Publications.

Hoffman, Griffin, G.A., Edwards, S.A., Paulissen, M. O., O'Neal, S.F., & Barnes, S.(1986). Teacher induction study: A final report of a descriptive study. Austin, TX: The University of Texas at Austin, Research and Developmental Center for Teacher Education.

House, R. (1980). Evaluating with validity. Beverly Hills, CA: Sage Publications.

Ilgen, D.R. & Youtz, M.A. (1986). Factors affecting the evaluation and development of minorities in organizations. In K. Rowland & C. Ferris (Eds.), Research in personnel and human resource management: A research annual (pp.307-337). Greenwich, CO:JAI Press

Irons, E., & Moore, G.W. (1985). Black managers in the banking industry. New York: Praeger.

Jones, E. (1986). Black managers: The dream deferred. Harvard Business Review, 64(3): 84-93.

Katims, D. & Henderson, R. (1990). Teacher Evaluation in Special Education. NASSP-Bulletin, 54(527), 47-52.

Kelley, M.F. (1985). The Arizona performance-based teacher certification program. Paper presented at the annual meeting of the American Educational Research Association, Chicago. (ERIC Document Reproduction Service No. ED 263078)

Kraiger, K., & Ford, J.K. (1985). A meta-analysis of ratee race effects in performance ratings. Journal of Applied Psychology, 70, 56-65.

Kraiger, K., & Ford, J.K. (1990). The relation of job knowledge, job performance, and supervisory ratings as a function of ratee race. Human performance, 3(4), 269-279.

Lane, B. A. (1990). Personnel evaluation: From problems to school improvement. Journal of research and development in education, 23(4), 243-249.

Lawrence, S., Biderman, M.D., Faley, R.H. (1987). An examination of employee perceptions of a subjective performance appraisal system. Journal of Business and Psychology, 2(2), 112-121.

Levenson, H., Buford, B., Bonno, B., & Davis, L. (1975). Are women still prejudiced against women? A replication and extension of Goldberg's study. Journal of Psychology 89, 67-71.

Lincoln, Y. & Guba, E. (1985). Naturalistic inquiry. Beverly Hills, CA: Sage.

Logan, C., Garland, J. & Ellett, C.D. (1989). Large-scale teacher performance assessment instruments: A synthesis of what they measure and a national survey of their influence on the preparation of teachers. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Louisiana State Department of Education. (1993). Louisiana assessor training manual: Intern teachers. (LDE R.S. 17:3721). Baton Rouge, LA: LDE Printing Office.

Louisiana State Department of Education. (1990). A statewide content verification study of teaching and learning components of the system for teaching and learning assessment and review (STAR). (Technical Report, No. 1). Baton Rouge, LA: LDE Printing Office.

Manning, R. (1988). The Teacher Evaluation Handbook. Englewood Cliffs, NJ: Prentice Hall.

Marczely & Bernadette. (1992). Teacher Evaluation: Research versus Practice. Journal of Personnel Evaluation in Education, 5(3), 279-90.

Martell, R.F., Lane, D., & Willis, S. (1992). A little sex bias can hurt women a lot. Paper presented at the Annual Meeting of the American Educational Research Association, Washington. (ERIC Document Reproduction Service No. ED 361398)

Martocchio, J. and Whitener, E. (1992). Fairness in personnel selection: A meta-analysis and policy implications. Human Relations, 45(5), 489-506.

McDonald, F. (1980). The problems of beginning teachers: A crisis in training (Vol 1). Study of induction programs for beginning teachers. Princeton, NJ: Education Testing Service.

McGreal, T. (1988). Evaluation for enhancing instruction: Linking teacher evaluation with staff development. In S. Stanley & J. Popham (Eds.), Teacher evaluation: Six prescriptions for success. Alexandria, Va: Association for Supervision and Curriculum Development.

McGreal, T. (1990). The use of rating scales in teacher evaluation: concerns and recommendations. Journal of Personnel Evaluation in Education, 4, 41-58.

Medley, D.M., & Coker, H. (1987). How valid are principals' judgements of teacher effectiveness? Phi Delta Kappan, 69(2), 138-140.

Mehrens, W., & Lehmann, I. (1991). Measurement and evaluation in education and psychology. Orlando: Holt, Rheinhart, and Winston, Inc.

Messick, S. (1989) Validity. In R.L. Linn (Ed.), Educational Measurement (3rd ed.) (pp.13-104). New York: American Council on Education. Macmillan.

Mount, M.K., & Ellis, R. (1987). Investigation of bias in job evaluation ratings of comparable worth study participants. Personnel Psychology, 40, 85-95.

Nelsen, E.A., & William, R.J. (1983). Observational ratings of teaching performance. Paper presented at the Annual Meeting of the American Psychological Association, Los Angeles.

Nixon, R. (1985a). Climbing the corporate ladder: Some perceptions among black managers. Washington, DC.: National Urban League.

Oosterhof, (1990). Classroom applications of educational measurement. Columbus: Merrill Publishing Co.

Park, B., & Rothbart, M. (1982). Perception of outgroup homogeneity and levels of social categorization: Memory for the subordinate attributes of in-group and out-group members. Journal of Personality and Social Psychology, 42, 1051-1068.

Parloff, M.B., Waskow, I.E., & Wolfe, B.E. (1978). Research on therapist variables in relation to process and outcome. In S. Garfield & A. Bergin (Eds.), Handbook of psychotherapy and behavior change. New York: Wiley. 233-282.

Patton, M.Q. (1990). Qualitative Evaluation and Research Methods (2nd ed.) Newbury Park, CA: Sage.

Peterson, Dehyle, & Watkins (1988). Minority teacher contributions. Urban Education.

Pheterson, G.I., Kiesler, S.B. & Goldberg, P.A. (1971). Evaluation of the performance of women as a function of their sex, achievement, and personal history. Journal of Personality and Social Psychology, 19, 114-118.

Price, M. (1989). Improving personnel evaluation processes: A synthesis of research and practice. Paper presented at the annual meeting of the Southern Regional Council on Educational Administration, Columbia.(ERIC Document Reproduction Service No. ED 313783)

Pulakos, E.D., White, E.A., Oppler, S.H., & Birman, W.C. (1989) Examination of race and sex effects on performance ratings. Journal of Applied Psychology, 74, 770-780.

Raymond, M. & Houston, W. (1990). Detecting and correcting for rater effects in performance assessment. Paper presented at the annual meeting of the American Educational Research Association, Boston, MA. (ERIC Document Reproduction Service No. ED 336429)

Remick, H. (1984). Major issues in a priori applications. In Remick, H (Ed.), Comparable worth and wage discrimination. Philadelphia, PA: Temple University Press. 99-117.

Rice, J., & Rice, D. (1973). Implications of the women's liberation movement for psychotherapy. American Journal of Psychiatry, 130, 191-196.

Robbins, T., & DeNisi, A. (1993). Moderators of sex bias in the performance appraisal process: A cognitive analysis. Journal of Management, 19(1), 113-126.

Rose, J.S. & Huynh, H. (1984). Technical issues for adopting the APT for districtwide teacher evaluation. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans. (ERIC Document Reproduction Service No. ED 247266)

Rosen, B., & Jerdee, T.H. (1973). The influence of sex role stereotypes on evaluations of male and female supervisory behavior. Journal of Applied Psychology, 57, 44-48.

Rosen, B., & Jerdee, T.H. (1974a). Influence of sex role stereotypes on personnel decisions. Journal of Applied Psychology, 59, 9-14.

Rosen, B., & Jerdee, T.H. (1974b). Effects of applicant's sex and difficulty of job on evaluations of candidate and managerial positions. Journal of Applied Psychology, 59, 511-512.

Ryan, K. (1979). Toward understanding the problem: At the threshold of the profession. In K. Hokey and R. Bents (Eds.), Toward meeting the needs of beginning teachers. Minneapolis, MN: United States Department of Education/Teacher Corps.

Sandefur, J.T. (1983). Competency assessment of teachers: 1980-1983. New York: Macmillian.

Schmidtt, N., & Hill, T.E. (1977). Sex and race composition of assessment center groups as a determinant of peer and assessor ratings. Journal of Applied Psychology, 62(3), 261-264.

Schwab, R.L. (1991). Research-Based Teacher Evaluation. Boston, MA: Kluwer Academic Publishers.

Scriven, M. (1988). Duty-based teacher evaluation. Journal of Personnel Evaluation in Education, 4, 41-58.

Sherman, J. (1980). Therapist attitudes and sex-role stereotyping. In A. Brodsky & R. Hare-Mustin (eds)., Women and psychotherapy: An assessment of research and practice. New York: Guilford. 35-66.

Soar, R.S., Medley, D.M., & Coker, H. (1983). Teacher evaluation: A critique of currently used methods. Phi Delta Kappan, 65(4), 239-246.

Spence, J.T., Helmreich, R., & Strapp, J. (1973). A short version of the attitudes toward women scale (AWS). Bulletin of the Psychonomic Society, 2, 219-220.

Stanley, S. & Popham, J. (Eds.), Teacher evaluation: Six prescriptions for success. Alexandria, VA: Association for Supervision and Curriculum Development.

Tanner, D. (1993). A nation 'truly' at risk. Phi Delta Kappan, 74 (4), 288-297.

Tawil, L., & Costello, C. (1983). The perceived competence of women in traditional and nontraditional fields as a function of sex-role orientation and age. Sex Roles, 9, 1197-1203

Taynor, J., & Deaux, K. (1975). Equity and perceived sex differences: role behavior as defined by the task, the mode, and the actor. Journal of Personality and Social Psychology, 32, 381-390.

Teddlie, C., & Roberts, S.P. (1992). Focus group results from the fall 1993 pilot study of the Louisiana Teacher Assessment Program for Interns (Interns and Assessors). Baton Rouge: Louisiana Department of Education.

Texas Education Agency. (1988). Teacher orientation manual 1988-1989. Austin, TX: Author

Tieman, C.R., & Rankin-Ullock. (1985). Student evaluations of teachers. Teaching Sociology 12, 177-191.

Tisher, R. (Ed.).(1978). The induction of beginning teachers in Australia. Melbourne, Australia: Monash University.

Treiman, DJ., & Hartmann, H. (1981). Women, work, and wages. Washington, DC: National Academy of Sciences.

Tyson, L.A., & Silverman, S. (1994). An analysis of physical education and non-physical education teachers at the elementary and secondary level on statewide teacher assessment. Journal of Teaching in Physical Education.

United States Department of Education. (1987). What's happening in teacher testing/ An analysis of state teacher testing practices. Washington, DC: Author.

Waldman, D.A., & Avolio, B.J. (1991). Race effects in performance evaluations: Controlling for ability, education, and experience. Journal of applied psychology, 76(6), 897-901.

Wall, T.C. (1984). The illusion of independence in evaluation. (Available from North Carolina State University)

Wise, A.E., Darling-Hammond, L., McLaughlin, M.W., & Bernstein, H.T. (1984). Teacher evaluation: A study of effective practices. Santa Monica, CA: Rand

Corporation. (ERIC Document Reproduction Service No. ED 246559)

Webster's Collegiate Dictionary (1988). Boston, MA: Houghton Mifflin Company

APPENDIX A
LOUISIANA TEACHER ASSESSMENT INSTRUMENT

DRAFT

LOUISIANA TEACHER ASSESSMENT INSTRUMENT

PRE-OBSERVATION CONFERENCE/INTERVIEW RECORD
INTERN TEACHER

Intern Teacher: _____ *Soc. Sec. No.:* ____ - ____ - ____

Assessor: _____ *Soc. Sec. No.:* ____ - ____ - ____

School District/Parish: _____ *School:* _____ *Date:* ___ ___ ___

KEY 1 = Needs improvement 2 = Proficient 3 = Commendable

Topic/Content: _____

Teacher presents written plan ☐ Yes ☐ No

Does this lesson cover new content or is it review/reteaching? _____

Domain I: Planning
Component A: The Teacher Plans Effectively for Instruction ☐

IA1 Specifies Learner Outcomes in Clear, Concise Objectives		1	2	3
Question Set 1 *What should the students know or be able to do at the end of this lesson and how did you arrive at your objectives? In your discussion, please indicate how this lesson relates to past and future lessons*	1 Nature of goals and objectives	•		•
	2 Selection of goals and objectives	•		•
	3 Planning beyond a single lesson	•		•

Comments/Notes:

KEY 1 = Needs Improvement 2 = Proficient 3 = Commendable

IA2 Includes Activity/Activities That Develop Objectives		1	2	3
Question Set 2 *What activities have you chosen to teach your objectives and why have you chosen these activities? In your discussion, please indicate the order in which I will see the activities.*	1. Selection of activities 2. Planned use of activities 3. Ordering of activities	-		-

Comments/Notes:

IA3 Identification and Plans for Individual Differences		1	2	3
Question Set 3 *Please identify individual differences you will address in this lesson and discuss how you plan to accommodate them.*	1. Identification of individual differences 2. Planned use for accommodation strategies/methods	-		-

Comments/Notes:

KEY 1 = Needs Improvement 2 = Proficient 3 = Commendable

IA4 Identifies Materials, Other Than Standard Classroom Materials, as Needed for Lesson		1	2	3
Question Set 4 *What media and materials other than textbook, workbook, or chalkboard will you use in this lesson, if any; and why have you chosen these resources?*	1. Selection of materials and media			
	2. Use of materials and media			
	3. Sequencing of materials and media in lesson			

Comments/Notes:

IA5 States Method(s) of Evaluation to Measure Learner Outcomes		1	2	3
Question Set 5 *Describe the method(s) you intend to use to measure learner outcomes of this lesson. When do you intend to use these methods?*	1. Identification of specific assessment methods			
	2. Appropriateness of assessment methods			
	3. Timeliness of assessment			

Comments/Notes:

KEY 1 = Needs Improvement 2 = Problem 3 = Commendable

Domain III: Instruction
Component D: The Teacher Assesses Student Progress

IIID1 Uses Assessment Technique(s) Effectively		1	2	3
Question Set 6 *How do you formally assess students' performance over the course of a unit, grading period, or year; and in what ways do you use the assessment results? Please be as specific as possible.*	1. Methods of assessment 2. Use of assessment data			

Comments/Notes:

IIID2 Monitors Ongoing Performance of Students		1	2	3
Question Set 7 *How do you monitor (that is, informally assess) performance of students; and how do you use the information? Please be as specific as possible.*	1. Methods of monitoring 2. Use of monitoring information			

Comments/Notes:

Note: Attribute ratings under Component D should be transferred to the Observation Record

KEY 1 = Needs Improvement 2 = Proficient 3 = Commendable

IIID3 Provides Timely Feedback to Students Regarding Their Progress		1	2	3
Question Set 8 *How and when do you give feedback to students about their performance? Please be as specific as possible.*	1. Use of feedback methods			
	2. Timeliness of feedback			

Comments/Notes:

9 Are there any special conditions of which I should be aware as I observe your lesson and classroom?

DRAFT

LOUISIANA TEACHER ASSESSMENT INSTRUMENT

OBSERVATION RECORD
INTERN TEACHER

Intern Teacher: _____ *Soc. Sec. No.:* ____ - ____ - ____

Assessor: _____ *Soc. Sec. No.:* ____ - ____ - ____

School District/Parish: _____ *School:* _____ *Date:* ___ ___ ___

Grade/Subject: _____ *Number of Students:* _____ *Class Hour:* _____

Observation No. 1 2 3 4 ☐ Support Semester ☐ Assessment Semester

SPECIAL CONDITIONS

☐ Special Populations ☐ Overcrowded Conditions ☐ Inadequate Facilities ☐ Inadequate Resources
☐ Other _____

Comments:

TEACHER OBSERVATION ANALYSIS AND SCORING SUMMARY

DOMAIN II: MANAGEMENT

Rating* Item*

☐ IIA MAINTA[redacted]ENT ☐ [redacted]

___ IIA1 Organizes to Facilitate Learning

___ IIA2 Promotes Positive Climate

___ IIB1 Manages Routines and Transitions

___ IIB2 Manages and/or Adjusts Time for Activities

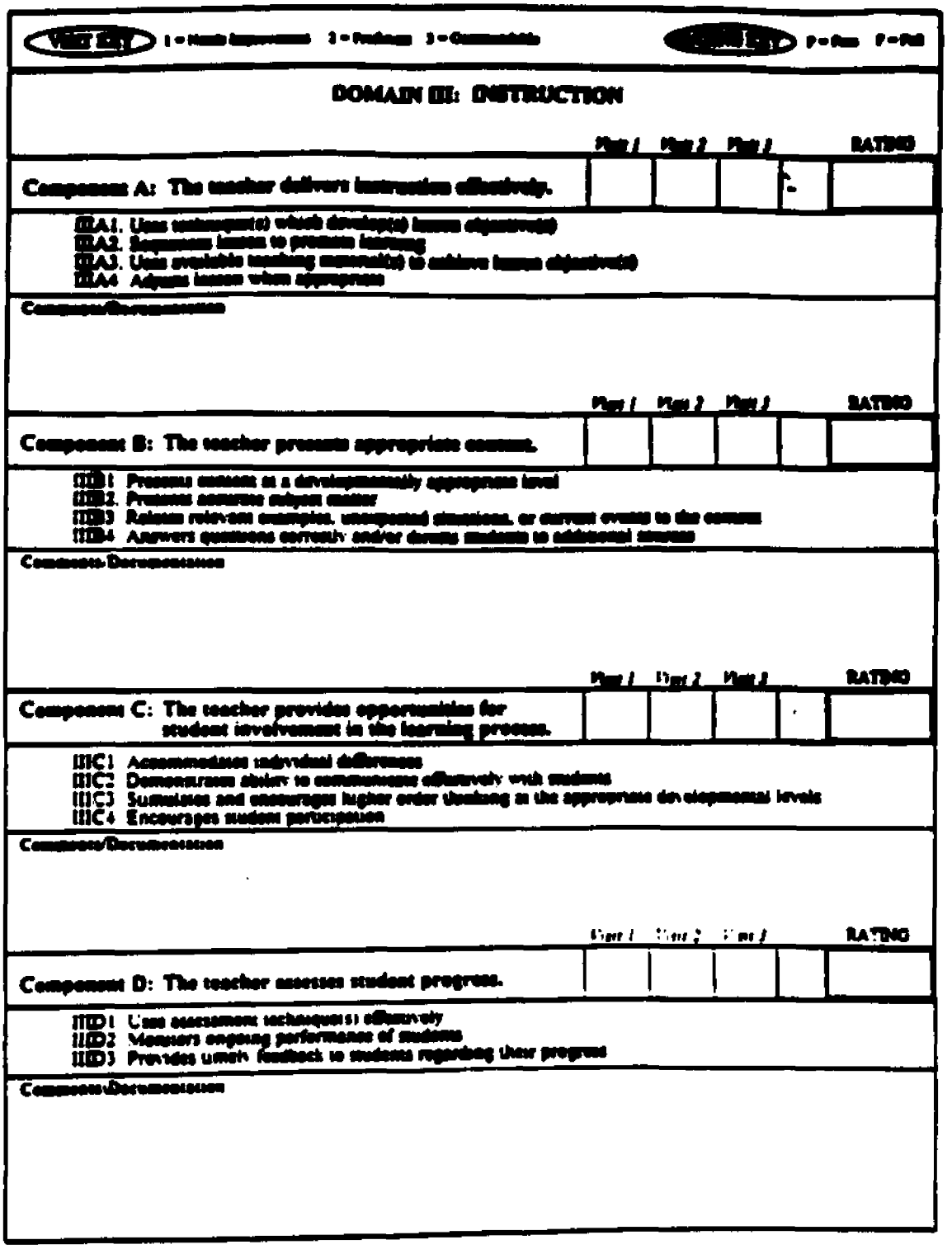
☐ IIC MANAGES LEARNER BEHAVIOR

___ IIC1 Establishes Expectations

___ IIC2 Uses Monitoring Techniques

* KEY 1 = Needs Improvement 2 = Proficient 3 = Commendable NO = Not Observed

Notes: All attributes may not be demonstrated in a lesson.
Supporting statements or referenced scripted lesson segments are required for each attribute rating.



DOMAIN III: INSTRUCTION

Component A: The teacher delivers instruction effectively. — Visit 1 | Visit 2 | Visit 3 | RATING

IIIA1. Uses assessment(s) which develop(s) lesson objective(s)
IIIA2. Sequences lessons to promote learning
IIIA3. Uses available teaching materials) to achieve lesson objective(s)
IIIA4 Adjusts lesson when appropriate

Comments/Documentation

Component B: The teacher presents appropriate content. — Visit 1 | Visit 2 | Visit 3 | RATING

IIIB1 Presents content at a developmentally appropriate level
IIIB2. Presents accurate subject matter
IIIB3 Relates relevant examples, unexpected situations, or current events to the content
IIIB4 Answers questions correctly and/or directs students to additional sources

Comments/Documentation

Component C: The teacher provides opportunities for student involvement in the learning process. — Visit 1 | Visit 2 | Visit 3 | RATING

IIIC1 Accommodates individual differences
IIIC2 Demonstrates ability to communicate effectively with students
IIIC3 Stimulates and encourages higher order thinking at the appropriate developmental levels
IIIC4 Encourages student participation

Comments/Documentation

Component D: The teacher assesses student progress. — Visit 1 | Visit 2 | Visit 3 | RATING

IIID1 Uses assessment techniques) effectively
IIID2 Measures ongoing performance of students
IIID3 Provides timely feedback to students regarding their progress

Comments/Documentation

Check if additional comments documentation are on back of page

DRAFT

LOUISIANA TEACHER ASSESSMENT INSTRUMENT

POST-OBSERVATION CONFERENCE RECORD
REGULAR CLASSROOM AND SPECIAL EDUCATION INTERN TEACHER

Intern Teacher: _____ *Soc. Sec. No.:* ____ - ____ - ____

Assessor: _____ *Soc. Sec. No.:* ____ - ____ - ____

School District Parish: _____ *School:* _____ *Date:* ___/___/___

1 Was the lesson consistent with information provided in the Pre-Observation Conference/Interview? Explain.

2 Strengths Exhibited:

Component/Attribute	*Comments/Commendations*

3 Areas for Improvement:

Component/Attribute	*Comments/Suggestions*

4 Teacher Comments: (optional)

5 Assessor Comments: (optional)

_____ _____

Teacher Signature/Date *Assessor Signature/Date*

Teacher's signature indicates that a Post-Observation Conference has been held and does not necessarily indicate agreement with the assessment comments.

DRAFT

LOUISIANA TEACHER ASSESSMENT INSTRUMENT

INTERN SUMMARY REPORT
INTERN TEACHER

ASSESSOR TEAM RATING FORM

Intern Teacher: _____ *Soc. Sec. No.:* ____ - ____ - ____

School District/Parish: _____ *School:* _____ *Date:* ___/___/___

VISIT KEY 1 = Needs Improvement 2 = Proficient 3 = Commendable RATING KEY P = Pass F = Fail

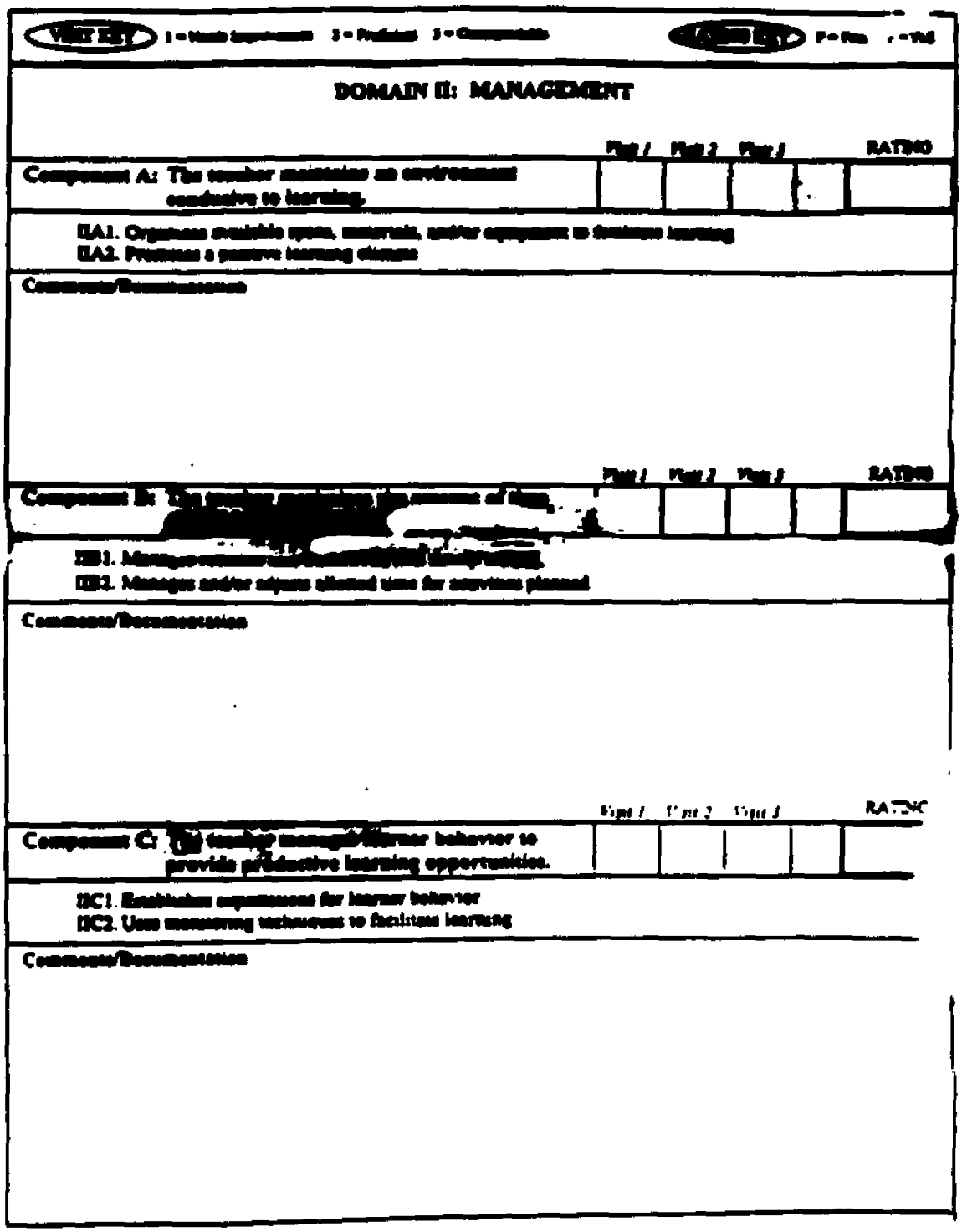
DOMAIN I: PLANNING

	Plan 1	Plan 2	Plan 3		
Competency #1: The teacher plans effectively for instruction.					

IA1. Specifies learner outcomes in clear, concise objectives
IA2. Includes activity/activities that develop objectives
IA3. Identifies and plans for individual differences
IA4. Identifies materials, other than standard classroom materials, as needed for lesson
IA5. States method(s) of evaluation to measure learner outcomes
IA6. Develops an Individual Education Plan (IEP), ITP, and/or IFSP
(Special Education Intern Teachers Only)

Comments/Documentation

— Check if additional comments
— Indicate where each visit was conducted



VISIT KEY 1 = [illegible] 2 = Proficient 3 = [illegible] RATING KEY P = Pass [illegible]

DOMAIN II: MANAGEMENT

	Visit 1	Visit 2	Visit 3		RATING
Component A: The teacher maintains an environment conducive to learning.					

IIA1. Organizes available space, materials, and/or equipment to facilitate learning
IIA2. Promotes a positive learning climate

Comments/Documentation

	Visit 1	Visit 2	Visit 3		RATING
Component B: [illegible]					

IIB1. [illegible]
IIB2. Manages and/or adjusts allotted time for activities planned

Comments/Documentation

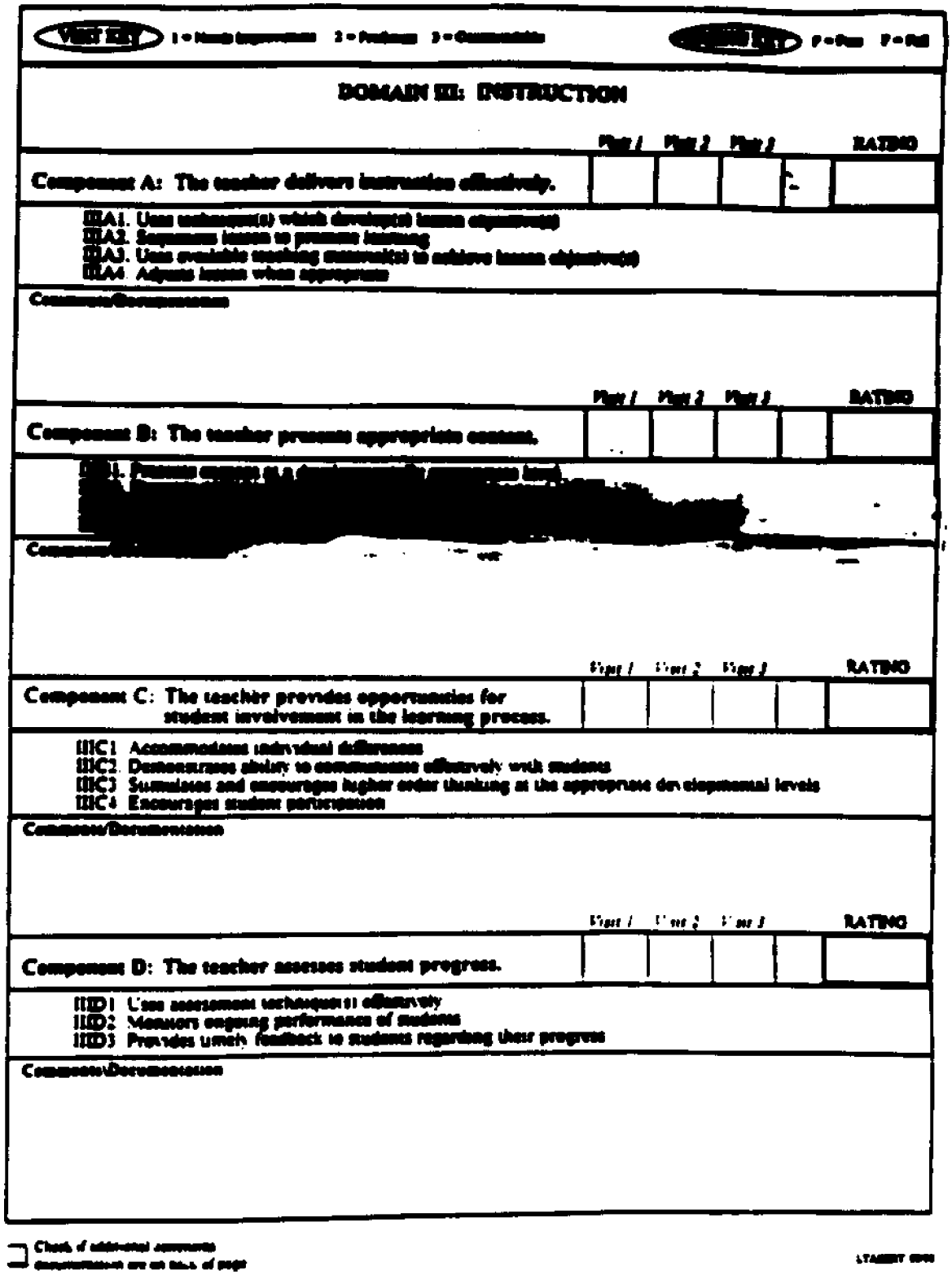
	Visit 1	Visit 2	Visit 3		RATING
Component C: The teacher manages learner behavior to provide productive learning opportunities.					

IIC1. Establishes expectations for learner behavior
IIC2. Uses monitoring techniques to facilitate learning

Comments/Documentation

Check if additional comments/documentation are on back of page

[illegible]



DOMAIN III: INSTRUCTION

Component A: The teacher delivers instruction effectively. — Visit 1 | Visit 2 | Visit 3 | RATING

IIIA1. Uses [illegible]
IIIA2. Sequences lessons to promote learning
IIIA3. Uses available teaching resource(s) to achieve lesson objective(s)
IIIA4. Adjusts lesson when appropriate

Comments/Documentation

Component B: The teacher presents appropriate content. — Visit 1 | Visit 2 | Visit 3 | RATING

[redacted]

Comments/Documentation

Component C: The teacher provides opportunities for student involvement in the learning process. — Visit 1 | Visit 2 | Visit 3 | RATING

IIIC1 Accommodates individual differences
IIIC2 Demonstrates ability to communicate effectively with students
IIIC3 Stimulates and encourages higher order thinking at the appropriate developmental levels
IIIC4 Encourages student participation

Comments/Documentation

Component D: The teacher assesses student progress. — Visit 1 | Visit 2 | Visit 3 | RATING

IIID1 Uses assessment techniques effectively
IIID2 Monitors ongoing performance of students
IIID3 Provides timely feedback to students regarding their progress

Comments/Documentation

☐ Check if additional comments [illegible] on back of page

APPENDIX B
LOUISIANA COMPONENTS OF EFFECTIVE TEACHING (LCET)

Louisiana Components of Effective Teaching Data Sources Intern Teachers		
Domains, Components, and Attributes		
DOMAIN I: PLANNING		
Component A: The teacher plans effectively for instruction.		
IA1. Specifies learner outcomes to clarify, focuses objectives		
IA2. Includes activities/exercises that develop objectives		
IA3. Identifies and plans for individual differences		
IA4. Identifies materials, other than standard classroom materials, as needed for lessons		
IA5. States method(s) of evaluation to measure learner outcomes		
IA6. Develops an Individual Education Plan (IEP), ITP, and/or IFSP (Special Education Intern Teachers Only)		
DOMAIN II: MANAGEMENT		
Component A: The teacher maintains an environment conducive to learning.		
IIA1. Organizes [illegible] space, materials, and/or equipment to facilitate learning		
IIA2. [illegible] learning centers		
Component B: The teacher maximizes the amount of time available for instruction.		
IIB1. Manages routines and transitions in a timely manner		
IIB2. Manages and/or adjusts allotted time for activities planned		
Component C: The teacher manages learner behavior to provide productive learning opportunities.		
IIC1. Establishes expectations for learner behavior		
IIC2. Uses monitoring techniques to facilitate learning		

Domains, Components, and Attributes		
DOMAIN III: INSTRUCTION		
Component A: The teacher delivers instruction effectively.		
IIIA1. Uses [illegible] which develop(s) lesson objective(s)		
IIIA2. Sequences lessons to promote learning		
IIIA3. Uses available teaching resource(s) to achieve lesson objective(s)		
IIIA4. Adjusts lessons when appropriate		
Component B: The teacher presents appropriate content.		
IIIB1. Presents content at a developmentally appropriate level		
IIIB2. Presents content in an [illegible] manner		
IIIB3. Relates relevant examples, unexpected responses, or current events to the content		
IIIB4. Answers questions correctly and/or directs students to additional sources (i.e., references, labs, learning centers, etc.)		
Component C: The teacher provides opportunities for student involvement in the learning process.		
IIIC1. Accommodates individual differences		
IIIC2. Demonstrates ability to communicate effectively with students		
IIIC3. Stimulates and encourages higher order thinking in the appropriate developmental areas		
IIIC4. Encourages student participation		
Component D: The teacher assesses student progress.		
IIID1. Uses assessment technique(s) effectively		
IIID2. Monitors ongoing performance of students		
IIID3. Provides timely feedback to students regarding their progress		

APPENDIX C
DEMOGRAPHIC DATA FORMS

LOUISIANA TEACHER ASSESSMENT PROGRAM
DEMOGRAPHIC DATA FOR ASSESSEES (Rev. 6/93)

Please print the following personal and professional information by responding completely to each item

NAME OF LOCAL SCHOOL DISTRICT _____

NAME OF LOCAL SCHOOL/UNIVERSITY _____

SOCIAL SECURITY # _____ TEACHING CERTIFICATE # _____

NAME _____
Last First Middle

Please supply the following information by filling in the appropriate response as shown (■)

Educational Levels *☐ BA/BS *☐ MA/MS *☐ MA/MS + 30
*☐ Ed.S *☐ Ph.D./Ed.D.

YEARS EXPERIENCE IN EDUCATION ☐☐

Gender *☐ Female *☐ Male

Current work areas (indicates the area in which you spend the most time teaching/working) You may fill in only one box

ASSESSEE TRAINER ☐YES ☐NO

*☐ Superintendent
*☐ Principal
*☐ Administrative Assistant
*☐ Assistant Principal
*☐ Counselor
*☐ Librarian
*☐ Speech Therapist
Special Education Teacher
*☐ Mild/Moderate
*☐ Severe/Profound
*☐ Non-Cat. Preschool
*☐ Academically Gifted
*☐ Sensory Impaired
*☐ Assessment Teacher
*☐ Resource Teacher

*☐ Lower Elementary Teacher (including Pre K and K)
*☐ Upper Elementary Teacher
*☐ Middle/Junior High Language Arts Teacher
*☐ Middle/Junior High Math or Science Teacher
*☐ Middle/Junior High Social Studies Teacher
*☐ Middle/Junior High Vocational Education Teacher
*☐ Secondary Language Arts Teacher
*☐ Secondary Math or Science Teacher
*☐ Secondary Social Studies Teacher
*☐ Secondary Vocational Education Teacher
*☐ Louisiana Department of Education

*☐ Foreign Language Teacher
*☐ Music Teacher
*☐ Art Teacher
*☐ Health or P.E. Teacher
*☐ Central Office Staff
*☐ University Faculty
*☐ Retired Teacher
*☐ Retired Administrator
*☐ Other (Please Specify)
_____

Please supply the following information by filling in the appropriate response as shown (■)

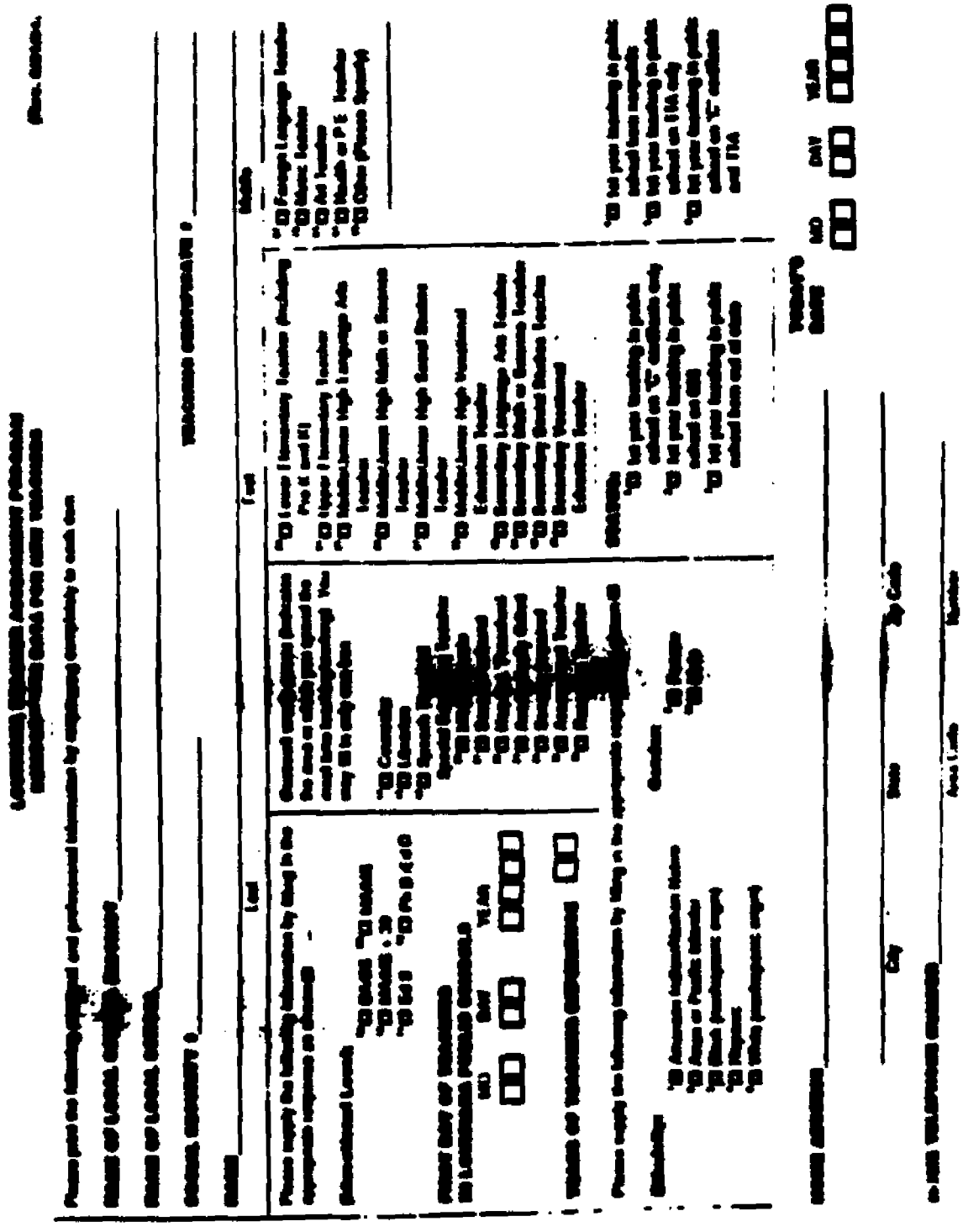
Ethnicity

'☐ American Indian/Alaskan Native
'☐ Asian or Pacific Islander
'☐ Black (nonhispanic origin)
'☐ Hispanic
'☐ White (nonhispanic origin)

NAME ASSESSED _____ TODAY'S DATE MO ☐☐ DAY ☐☐ YEAR ☐☐☐

_____
City State Zip Code

HOME TELEPHONE NUMBER _____ _____
Area Code Number



APPENDIX D
SURVEY OF EFFECTIVE TEACHING BEHAVIOR

Dear Educator:

My name is Shana Schuyten and I am a doctoral student in Educational Research Methodology at Louisiana State University. I am currently working on the collection of my dissertation research under the guidance of Dr. Abbas Tashakkori. My research addresses the perceptions and attitudes of effective teaching behaviors and practices. As an educator, your input is desperately needed and essential to this area of study. The results of this research will be beneficial in furthering our understanding of how instructional and administrative personnel perceive effective teaching strategies.

Several weeks ago I wrote to your school superintendent requesting permission to survey a small sample of experienced teachers, principals and central office employees within your district. Having been given permission to conduct my dissertation research in your district, I am now requesting your participation which will take only a few minutes of your time.

Enclosed is a two page survey instrument developed to study the perceptions and attitudes of effective teaching behaviors and practices. The survey will take you less than 15 minutes to complete and would ideally be completed during your preparation period. The instrument consists of a detailed description of a classroom teaching scenario and a response form. After you have completed reading the classroom teaching scenario on page one you are asked to respond on page two by rating the teacher's perceived level of effectiveness. For purposes of this survey, an effective teacher is one whose teaching attitudes, strategies, practices and behaviors result in positive learning outcomes and successes. The names of all districts, schools, principals, and teachers who participate in the survey will remain confidential. No one but the researcher will know which schools and districts participated in the study. The results of the effectiveness survey will also remain confidential.

By understanding the rationales behind teacher behavior in the classroom environment, the ultimate goal of utilizing effective teaching strategies will be met. Therefore, I am requesting your cooperation in filling out this brief survey. Please return only the response page (page 2) of the survey by February 18, 1995 using the self-addressed stamped envelope that I have enclosed for your convenience.

Your cooperation is greatly appreciated and your input is vital. I thank you in advance for your participation and look forward to hearing from you.

Shana N. Schuyten

Shana N. Schuyten
1773-A Boulevard De Province
Baton Rouge, LA 70816
(504) 273-0880

District Permission Form for Research Project

Permission is given for Shana N. Schuyten to conduct a survey of experienced teachers, principals, and central office personnel in selected schools within our district. The purpose of this research is to fulfill her dissertation requirements. All participants' responses will be confidential and participant and district information will not be identified as to any specific district. Twenty districts across the state of Louisiana will participate in this research effort. The researcher, Shana N. Schuyten, will be the only one with access to individual surveys, which will be anonymous at the participant level.

Please return by January 24, 1998

Public School District _____

Superintendent's Signature _____

Date: _____

NOTE: **All responses will be confidential at the participant, school and district level. No one or organization will have access to the individual district, school or participant information except the researcher.**

Any description given of this study should be very broad so as not to influence participant's responses to the survey.

If you have any questions or concerns, please call Shana N. Schuyten at (504) 273-0880.

Please mail to : Shana N. Schuyten
1773-A Boulevard De Province
Baton Rouge, LA 70816

Mrs. Jones is a white science teacher with 5 years of teaching experience at Johnson Elementary School. She teaches average fourth grade students and has a class size of 22 students. Mrs. Jones is teaching a two week unit on Louisiana Wildlife with today's lesson focusing on the crawfish, its anatomy, its origin and its economic impact on the state. Mrs. Jones had asked her students the day before to draw a picture of what they thought a crawfish looked like and on the bottom of the page to write down three things they knew to be true about the crawfish. The students would share their answers following the lesson.

[redacted]

Mrs. Jones's classroom has been arranged with pairs of desks facing each other. Near the front of the classroom is a clear glass tank full of live crawfish and a crawfish diagram has been drawn on the chalkboard. The children immediately begin to crowd around the tank, poking and prodding at the crawfish and some children begin to push each other hoping to get closer to the tank. Mrs. Jones is near the back of the classroom and notices the students crowding the tank. She immediately asks them to report to their assigned desks. Mrs. Jones then reminds the children of the classroom rules and appropriate conduct.

[redacted]

Mrs. Jones begins her lesson with a brief overview of what the class will be doing in the next 55 minutes and reminds the children to listen carefully. As she speaks, several children fidget in their desk, talking among themselves but Mrs. Jones continues. Mrs. Jones tells the class that they are going to spend time watching crawfish, collecting some information on them, and discussing how important the resource is to Louisiana. Mrs. Jones has clearly researched the crawfish and is very knowledgeable about the lesson content. Mrs. Jones knows of the crawfish's origin, its anatomy, and it contributions to Louisiana. However, as she continues with the lesson, she discusses the catching, cooking and preparing of food dishes made with crawfish and never really addresses how the crawfish is a valuable resource impacting Louisiana.

Next, Mrs. Jones gives instruction on how to handle the crawfish and asks the students to handle the crawfish with care, reminding them that they are live creatures, and to only handle them when necessary. Mrs. Jones then demonstrates how exactly to handle the crawfish, walking around the classroom so that every student can see clearly. The students appear very curious and begin to make clever remarks about the crawfish , and some try reaching out to touch the crawfish in Mrs. Jones' hand. Mrs. Jones calls one child from each of the 11 teams to come to the front of the room to get a crawfish and gives them each a box to transfer the crawfish, while the remaining students sit with nothing to do. Once the students are back at their desks with the crawfish, she attempts to hand out the materials for the lesson which consist of a ruler, magnifying glass and a crawfish diagram. The magnifying glasses have been misplaced and it takes a few moments for her to locate them. There are enough rulers and crawfish diagrams for every child. Having gotten their crawfish and materials, Mrs. Jones asks the children to pay attention to the tasks at hand although some students are distracted by the crawfish trying to climb out of the box. Mrs. Jones moves around the room and questions the students about what they are seeing and feeling. She asks students to point out various parts of the crawfish and has some of them examine the crawfish more carefully to locate correct answers when they make a mistake. With ten minutes left in the class period, all students are asked to stop what they are doing and to write a paragraph about what they have learned about the crawfish. A few students continue to handle the crawfish , ignoring Mrs. Jones' instructions. They comply when they are individually instructed to put them down by Mrs. Jones. The recess bell rings before she is able to collect the papers, materials and the crawfish .

1) In your opinion, How Effective is Mrs. Jones' Teaching?

1 2 3 4

Not Effective Highly Effective

2) What do you consider to be Mrs. Jones' Most Effective Teaching Behavior?

3) What do you consider to be Mrs. Jones' Least Effective Teaching Behavior?

4) What might she have done to have taught more effectively?

For purposes of collecting demographic information, please respond to the following:

5) Type of school you currently teach in:
Elementary___ Middle/ Jr. High School___ High School___ Combination___

6) Primary subject (s) you teach:_____

7) Total years of teaching experience:_____

8) Sex: Male___ Female___

9) Race: Caucasian___ African American___ Hispanic___ Asian___ Other___

10) Age: 20-25___ 26-30___ 31-35___ 36-40___ 41-45___
46-50___ 51-55___ 55 +___

VITA

Mrs. Shana Lewitt Schuyten was born on November 13, 1969, in Nashville, Tennessee, the only daughter of Allan Jacob Lewitt and Sharon Ellett Lewitt. After graduating from Robert E. Lee High School (Baton Rouge, Louisiana.) in 1987, she accepted an academic scholarship to attend Newcomb College at Tulane University, where she was a member of the varsity swim team. Four years later, Shana earned a B.A. degree with a major in psychology.

Shana's next academic endeavor began only three months later when she enrolled at Louisiana State University to pursue a doctoral degree in Educational Research Methodology within the College of Education. Shana was the recipient of a graduate teaching and research assistantship for three years during her pursuit.

Upon completing the majority of her coursework, she accepted an appointment to the Louisiana Department of Education as a Psychometrician in the Bureau of Professional Accountability, where she was employed for a year and a half. Upon leaving the Department of Education, Shana gained employment as an Educational Specialist for SERVE! Mid City, a federally and locally funded Americorps program serving the children of the Baton Rouge Mid City Area. Shana remains employed as an Educational Specialist.

After four long and trying years of personal commitment, sacrifice, and dedicated study, Shana will be receiving her

long-awaited degree in August of 1995. Shana will have finally and successfully attained her lifelong goal.

DOCTORAL EXAMINATION AND DISSERTATION REPORT

Candidate: Shana Lewitt Schuyten

Major Field: Educational Research

Title of Dissertation: The Effects of Assessor and Assessee Gender, Ethnicity, and Assessor's Role on Performance Assessment of Teachers

Approved:

Major Professor and Chairman

Dean of the Graduate School

EXAMINING COMMITTEE:

Date of Examination: 6/14/95